12 Probability and Genetics

Hutz: Ladies and gentlemen, I'm going to prove to you not only that Freddy Quimby is guilty, but that he is also innocent of not being guilty. I refer you to my expert witness, Dr. Hibbert.
Hibbert: Well, only one in two million people has what we call the "evil gene". Hitler had it, Walt Disney had it, and Freddy Quimby has it.
Hutz: Thank you, Dr. Hibbert. I rest my case

From: *The Simpsons*

12.1 Basic Terms

We assume the reader is familiar with the basic ideas of genetics as proposed by G. Mendel in the 19th century. The traits of an organism (its phenotype) are controlled (at least in part) by factors called genes, which can be passed from the parents to the offspring through sexual reproduction. Mendel proposed this mechanism as an explanation for the observed pattern of phenotypes that he obtained in numerous systematic crossings of pea plants. The physical existence of genes and the explanation of how they function occurred much later. We recall some basic vocabulary in the subject as it is presented today.

Gene: A portion of DNA that encodes for a protein. The human genome consists of approximately 30,000 genes.

Chromosome: The physical unit in the cell that contains the genes and which is passed to the offspring as the basic unit of inheritance, possibly with some alteration. In diploid cells the chromosomes are present in pairs, called *homologous* chromosomes. (46 chromosomes or 23 pairs in humans)

Locus: The physical portion of a chromosome that contains a particular gene.

Allele: A variant form of a gene, designated $A_1, A_2, \ldots A_n$ if there are $n \geq 3$ forms, but A, a if there are only two forms and A if there is only one.

Gamete: The male and female sex cells. These contain one chromosome from each homologous pair and thus have half the number of chromosomes of a somatic cell (referred to as haploid cells.)

Zygote: A fertilized cell obtained from joining a male and female gamete. This cell contains the diploid or full number of chromosome pairs.

Genotype: The pair of alleles of a given gene or genes that appear at a particular locus. The genotype describes the possible gametes which can be formed by the adult, ignoring mutations and crossovers.

Phenotype: The actual physical characteristic produced by a genotype.

We will apply the counting methods we have developed and the basic notions of probability to some typical problems in the genetics of individuals and population. These results can be found in most books on genetics or population ecology. Our intent is to emphasize the mathematical unity behind the solution of these problems.

12.2 Counting Problems

We first consider problems of enumerating the genotypes. The physical relationship between genes and chromosomes complicates the counting. The essential biological point to bear in mind is the following principle:

Principle of Equivalence: Interchanging all the alleles from one chromosome in a homologous pair to the other doesn't change the possible gametes that the adult can produce with respect to the given genes. Hence, we must count these as one genotype.

Most of the enumeration problems one encounters in genetics can be handled using the multiplication principle of counting, with the results adjusted to take into account the Principle of Equivalence. This is exhibited immediately in the first example.

Example 12.1: List the possible genotypes if there are two alleles A and a at a particular locus.

Solution:

Suppose the pair of homologous chromosomes is #1 and #2. There are two choices for the allele on chromosome #1, namely A and a. Each of these can be paired with allele A or a on chromosome #2. The multiplication rule implies that there are $2\times 2 = 4$ ways to make the two assignments, namely AA, Aa, aA and aa. But the two selections Aa and aA arise from switching the allele assignments. According to the Principle of Equivalence these two arrangements count as a single genotype. Thus there are only three genotypes, AA, aA, and aa. ■

Example 12.2: List the possible genotypes if there are three alleles A_1, A_2, A_3 at a locus.

*Solution***:**

As in Example 12.1 the multiplication principle implies that there are nine ways of assigning the two alleles to the chromosome pair. However, this counts as distinct choices, for example, A_1A_2 and A_2A_1, which according to the Principle of Equivalence describe the same genotype. There are 6 mixed assignments of this sort and we must only count half of them; the other three should be discarded. Thus, from the original count of 9 we are left with 6 distinct genotypes. These are the three homogeneous types (*homozygotes*) A_1A_1, A_2A_2, A_3A_3 and three heterogeneous types (*heterozygotes*) A_1A_2, A_1A_3, A_2A_3. ■

We often want to categorize the genotype with respect to two different loci. The alleles at one locus will be designated $A_1, A_2, \ldots$ (or A and a if only two alleles), while the alleles at the other locus will be designated $B_1, B_2, \ldots$ (or B and b if only two alleles). In this case we will see that the number of genotypes depends on whether the loci are part of the same homologous pair or on different pairs.

Example 12.3: If there are two alleles at each of two loci on different pairs of chromosomes, how many genotypes are possible with respect to the two loci?

Solution:

From Example 12.1 we can assign three genotypes to locus 1, namely AA, Aa, and aa. These can be paired with the three possible genotypes BB, Bb, and bb available at locus 2. The multiplication principle implies that with respect to the two loci there are 3×3 or nine genotypes. We denote these, for example, by $Aa\,//\,Bb$, where the entry before the double slash indicates the genotype on the first locus (which only involves the A gene), and the second term indicates the genotype on the second locus (which only involves the B gene). The double slash denotes that the genes are on different chromosome pairs.■

Example 12.4: If there are two alleles at two different loci on the same chromosome pair, how many different genotypes are possible?

Solution:

On each homologous chromosome in the pair there are four possible arrangements at the two loci: AB, Ab, aB, ab. The second chromosome in the pair can also receive one of these four arrangements. The multiplication principle gives 16 ways of assigning the alleles to the two chromosomes. However, we have to take into account the Principle of Equivalence. For example, using a single / to separate the two chromosomes in a homologous pair, the combination $Ab\,/\,aB$ describes the same genotype as $aB\,/\,Ab$. How many such redundancies are there?

A redundancy will occur when the arrangement involves different choices for the two chromosomes. We can select the assignment for the first chromosome in any of four ways, but we then have only three selections available for the second chromosome, if we want the two selections to differ. The multiplication principle implies that there are 12 arrangements of this sort, and by the Principle of Equivalence, half of them are redundant. Thus we are left with $16-6=10$ genotypes.

It seems perhaps paradoxical that there are more genotypes in the case considered here than in the previous example. However, taking into account the second locus on the chromosome effectively makes the two homologous chromosomes somewhat more distinguishable from each other and leads to the greater number of possibilities.■

As the previous examples show, the number of genotypes grows very quickly with even a small number of modestly polymorphic (i.e. more than one allele) genes. Because of this, in many applications of population genetics, for instance forensic blood type matching, it is customary to deal directly with the genes and their frequencies, rather than the genotypes. The Hardy-Weinberg theorem (discussed below, Theorem 12.2) allows one, under certain circumstances, to use the information on genes to calculate the information on genotypes, which is usually the ultimate objective.

If the phenotype of Aa is the same as the phenotype AA we say that the allele A is *dominant* and a is *recessive*. In the case of two alleles, we shall always denote the dominant one, if there is one, by A. If no dominance pattern is stated the three genotypes give three distinct phenotypes.

Example 12.5: If A and B are each dominant for different traits and the loci of A and B lie on separate chromosomes (i.e. the loci are not on a homologous pair) how many phenotypes correspond to the nine genotypes described in Example 12.3.

Solution:

There are two phenotypes associated with each gene, namely A, a and B, b. The multiplication principle implies that there are 4 possible phenotypes. This can also be obtained by examining each of the nine genotypes and classifying them according to their phenotype. For example, $Aa\,//\,bb$ would give the phenotype Ab.■

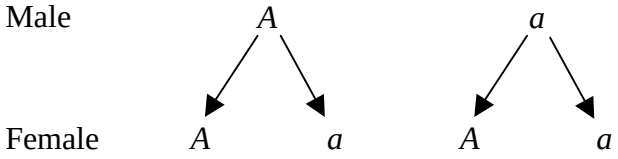
12.3 Reproduction of the Organism

We assume the organism reproduces sexually. If we know the genotypes of the parents, then Mendel's theory allows us to calculate the probability, and therefore the expected relative frequency for the genotype of the offspring.

Example 12.6: A male of genotype Aa at a particular locus is crossed with a female with the same genotype at that locus. List all possible genotypes of the zygotes and the probability of each.

Solution:

The male produces gametes with genes A and a in equal number, similarly for the female. A random zygote is obtained by selecting a male gamete and a female gamete independently of each other. There are four possible selections, as a tree diagram reveals.



Using the independence of the choices we have the following probabilities: (the first letter denotes the gene selected for the male gamete.)

$$P(AA)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4},\ P(aA)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4},\ P(Aa)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4},\text{ and }\ P(aa)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}.$$

The two middle outcomes give the same genotype for the zygote. Therefore the probabilities for the genotype frequencies for the zygotes are $AA:1/4,\quad Aa:1/2,\quad aa:1/4$ ■

Example 12.7: Suppose that with respect to two genes on non-homologous chromosomes the parents are both of type $Aa//Bb$. What is the probability of occurrence for each of the possible genotypes for the zygotes? (This type of pairing is referred to in genetics as a dihybrid cross.)

Solution:

When the gametes are formed by meiosis, the chromosomes from each pair are sorted independently i.e. the genotype created for one locus is independent of the genotype created at the other locus (Mendel's principle of independent assortment). Therefore, for each genotype of a zygote, for example $Aa//bb$, its probability of occurrence is obtained by multiplying the probability of obtaining each separate genotype, which was given in Example 12.6. Thus $P(Aa//BB)=\frac{1}{2}\times\frac{1}{4}=\frac{1}{8}$. Going through the nine possible genotypes of a zygote gives

AA//BB	1/16	*Aa//BB*	1/8	*aa//BB*	1/16
AA//Bb	1/8	*Aa//Bb*	1/4	*aa//Bb*	1/8
AA//bb	1/16	*Aa//bb*	1/8	*aa//bb*	1/16

■

If the two loci of genes A and B are on the same chromosome, the analysis of the pairings of males and female is more difficult. Mendel's principle of independent assortment no longer applies to the production of gametes. Indeed, during the meiotic division leading to the creation of the gametes, the phenomenon of crossover may lead to new combinations of genes in the gametes. Specifically, if the genotype is heterozygous for each allele, for example AB/ab, then when gametes are formed, the alleles B and b may be exchanged, yielding recombinant gametes Ab and aB, in addition to the so-called parental gametes AB and ab. A similar result applies when the parental genotype is Ab/aB.

In this situation the gametes are produced as if the genes were on separate chromosomes, except that the four gametes types AB, Ab, aB, and ab may not occur with equal frequency. We denote by r the probability that a gamete is of recombinant type (also called the *recombination frequency*) and by p the probability that it is of parental type. Mathematically, there are two aspects of the crossover mechanism that are important. First, when there is crossover the four haploid cells produced by the adult contain one of each gamete type. (Note that meiosis produces

four gametes, rather than two, because the chromosomes double before the meiotic division.) This implies that there is an equal chance of occurrence ($r/2$) of each of the two recombinant types. Moreover, when no crossover occurs only gametes of parental type are produced and in equal numbers. Thus each parental genotype has probability of occurrence equal to $p/2$. Since crossover produces both parental types and recombinant types in equal numbers and non-crossover produces only parental types, we must have $p \geq r$. We can now deduce the following fundamental result.

Theorem 12.1 (Recombination frequency):

In a doubly heterozygous adult the frequency r of recombinant gametes and the frequency p of parental type gametes satisfy

a) $p + r = 1$

b) $0 \leq r \leq \frac{1}{2}$.

Moreover, each recombinant genotype occurs with probability $r/2$ and each parental genotype with probability $p/2$.

Solution:

Since any gamete either is of recombinant or parental type a) follows from the definitions of r and p. From the remarks preceding the theorem, we have that $p \geq r$ and therefore $1 = p + r \geq 2r$ from which b) follows. The last statement summarizes the discussion of the crossover mechanism given above.■

Note that if $r = 1/2$ then also $p = 1/2$ and all four gametes are produced in equal proportions, as if the genes A and B were on different chromosomes. This tends to occur when the loci are far apart on the chromosome. When $r = 0$ or is very small, crossover is rare and the parental genotypes predominate in the gamete pool. In this case the gene loci are near each other. The recombination frequency r is important because it provides a measure of the "distance" between loci on the same chromosome. This frequency can be often be estimated using testcrosses with pure-breeding recessives, as described in exercise 10.

12.4 Reproduction of the Population

We have seen above how to predict and assign probabilities to the possible outcomes for the mating of individuals. Similar techniques can be used to describe the probable development for a large ensemble of individuals, known as a population. We focus first on a particular gene with two alleles. What can be said about the frequencies of the three genotypes in the population as a whole? The answer, known as the Hardy-Weinberg principle is surprisingly simple. However, it requires certain assumptions on the population and the mating of individuals. These assumptions are

(i) The population is large. This enables us to think of relative frequencies as probabilities and vice-versa.

(ii) The mating is random. The genotype in question has no effect on an individual's choice of mate. We can think of this figuratively as if the males were placed in one large urn, the females in another and fate picks partners from each urn at random and independently.

(iii) Males and females have the same distribution of the three genotypes at the given locus. In particular, the locus cannot be on the so-called X chromosome, of which males only carry one homologous copy. The analysis of sex-linked traits is more complicated, although a similar result to the Hardy-Weinberg principle is true. See exercise 14 for the details.

(iv) Each genotype has equal fitness. This means that all zygotes have an equal chance of surviving to reproductive age.

(v) There are no mutations at the locus or migrations of individuals into the population.

Our analysis of the genotype frequencies for populations uses the mathematical artifice of the gene pool. We think of the genes from all males as being in one large collection and similarly the genes of all females. A mating can be viewed as simply picking a member from each of these pools to create an offspring. The schema in all these analyses can be represented as follows:

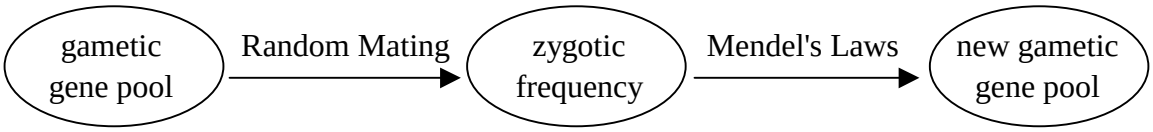


Figure 12.1

The assumption of random mating (and the laws of probability discussed in the previous chapters) enables us to derive the zygote frequency given some initial distribution of genes. Mendelian genetics then allows us to predict the frequency of the genotypes in the next generation. We first illustrate this procedure in its simplest case, the Hardy-Weinberg equilibrium theorem.

Suppose the allele A has frequency p and the allele a has frequency q, where $p+q=1$. It is convenient to express this information (the first "bubble" in Figure 12.1) in tabular form, as this helps with the analysis of more complicated examples.

Initial Gametic Frequency (Male & Female)		
Gamete Type	A	a
Frequency	p	q

By (iii) these frequencies are the same in the male and female gene pools. Interpreting the frequencies as probabilities, (i), we can find the frequency of each of the three genotypes, AA, aA, aa. For example, the genotype AA arises from the selection of an A allele from the

male gene pool and an A allele from the female gene pool. As these are independent events, (ii), the probability of obtaining an offspring of genotype AA is the product of the two relative frequencies $p \times p = p^2$. Similarly, the probability of obtaining the genotype aa is q^2, while the probability of obtaining the genotype aA is $2pq$ since we can create this genotype by the selection of an a male and an A female, or vice-versa. This gives us a table for the second step in Figure 12.1.

Zygote Frequency			
Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2

To summarize: If we begin with a large population having gene frequencies for the two alleles of p for A and q for a, then the zygotes will be created in the proportions $AA : p^2$, $aA : 2pq$, $aa : q^2$. These three ratios have a remarkable property. The adults in this generation will produce gametes with A and a in the exactly the same proportions as the parental generation. Let's see why this is so.

The zygotes described in the preceding paragraph mature and according to (iv) and (v) the genotypes of the adults will have the same proportions. What will be the gene pool frequencies, denoted by p' and q', for these new adults? (This is the third "bubble" in Figure 12.1.) We can compute these frequencies using the rules of probability developed in Chapters 10 and 11.

The selection of a gene from the new pool can be visualized as a two-stage process. First we select an adult and then we select one of that adult's genes. For example, the event that an A gene is selected from the pool is achieved in one of two mutually exclusive ways. Either we first select an adult of genotype AA and then select an A gene from its gametes or we select an adult of type Aa and then select its A gene. (Selecting an adult of type aa is obviously of no value here.) Using a slightly ambiguous but simple notation, this decomposition yields the probability formula

$$p' = P(A) = P(AA \text{ and } A) + P(Aa \text{ and } A).$$

The joint probabilities on the right may be evaluated using conditional probabilities. We have

$$P(AA \text{ and } A) = P(A \mid AA)P(AA) = P(AA)$$

and

$$P(Aa \text{ and } A) = P(A \mid Aa)P(Aa) = \tfrac{1}{2}P(Aa).$$

This yields the following useful formula expressing the gene frequency of A in terms of the relevant genotype frequencies:

$$P(A) = P(AA) + \tfrac{1}{2}P(Aa) \qquad (12.1)$$

Using for $P(AA)$ and $P(Aa)$ the zygote frequencies listed above, formula (12.1) yields that

$$p' = P(A) = p^2 + pq = p(p+q) = p$$

since $p+q=1$. Similarly the reader should show that $q' = P(a) = q$. Thus the table for the new gametic frequencies is identical to the original distribution.

New Gametic Frequency (Male & Female)		
Gamete Type	A	a
Frequency	$p' = p$	$q' = q$

What do these results mean? No matter the gene frequencies we started with, (the p and q), if selection is not operative and mating is random, then after one generation the gene frequencies will be precisely the same as in the initial state. Since in random mating the gene frequencies determine the genotype frequencies for the offspring, we obtain the Theorem of Hardy & Weinberg:

Theorem 12.2 (Hardy-Weinberg Equilibrium): If assumptions (i) to (v) above are true, then after one generation the genotype frequencies will have the ratios $AA:p^2$, $aA:2pq$, $aa:q^2$. These ratios will persist as long as assumptions (i) to (v) remain in effect.■

A population is said to be in Hardy-Weinberg equilibrium if the genotype ratios have the values given in the theorem. These ratios give a baseline measure for determining whether the five conditions above are operative. Deviations from the Hardy-Weinberg ratios indicate that at least one of the five assumptions is not true. Typically, the mating may not be random or a selection pressure may be in effect that decreases the fecundity of one genotype compared with the others.

Example 12.8: The MN blood grouping is a blood classification based on a single gene with two alleles, M and N. Each of the three genotypes MM, MN, and NN produces a distinct phenotype, which are usually designated M, MN and N, respectively. In a sample of 6129 European-Americans the observed numbers with these phenotypes (and hence the corresponding genotypes) were found to be:

Genotype	MM	MN	NN
Frequency	1787	3039	1303

Based on this data, is the population in Hardy-Weinberg equilibrium? (Data from *Evolution*, 2nd Edition by Mark Ridley, Blackwell Science Publ.)

Solution:

We must compute the gene frequencies $p = p_M$ and $q = q_N$ based on this sample.

- First find the frequencies of the three genotypes. These are $MM: \frac{1787}{6129} \approx .292$, $MN: \frac{3039}{6129} \approx .496$, and $NN: \frac{1303}{6129} \approx .213$.

- Next we must compute the gene frequencies using these genotype frequencies. To do this use equation (12.1) above. We find that the frequency of M is $p = .292 + \frac{1}{2} \times .496 = .54$ and the frequency of N is $q = .213 + \frac{1}{2} \times .496 = .46$. Note that $p + q = 1$, as required for the two gene frequencies.

- We now compute the Hardy-Weinberg ratios using the gene frequencies just found. The predicted genotype frequencies are *MM*: $p^2 = (.54)^2 = .292$, *MN*: $2pq = 2 \times .54 \times .46 = .497$, and *NN*: $q^2 = (.46)^2 = .212$. These are very close to the observed frequencies, indicating that the population is in H-W equilibrium at this locus. More ambiguous cases can be evaluated using a statistical technique known as the chi-squared test.■

The Hardy-Weinberg theorem extends easily to genes that have more than two alleles. (See for example exercise 12.) When an equilibrium exists it is permissible to compute genotype frequencies in the population by multiplying the respective gene frequencies. This is quite useful for example, in forensic DNA identification where highly polymorphic genes are often examined and it would be extremely time-consuming and expensive to obtain statistical data on the large number of possible genotypes. Of course the genes that are used must satisfy the hypotheses (i) to (v) listed earlier. In particular, the migration issue (v) requires a definition of the relevant breeding population and this has led to some controversies when applied to a society such as the U.S. with large subpopulations.

One might expect that a result similar to the Hardy-Weinberg theorem would hold for genotype frequencies at multiple loci. For example, if the genes A and B, each with two alleles, were located on different homologous pairs one might hope that any initial distribution of the four possible gametes *AB, Ab, aB, ab* would lead to a stable genotype distribution. This is indeed the case, but it takes more than one generation to establish. For example, suppose we began with only two gamete types AB: frequency 0.5 and ab: frequency 0.5 in each of the male and female gamete pools. The formation of a zygote involves the selection of a gamete from the male gene pool and one from the female. The zygote $Aa // Bb$ can occur either by selecting an AB gamete from the male pool and an ab gamete from the female pool or vice-versa. Therefore, assuming these choices are made independently, the zygote genotype $Aa // Bb$ will occur with frequency $2 \times .5 \times .5 = .5$. The gametes produced by this zygote contain type Ab, which did not exist in the original gene pool. In fact, the frequency of this gamete, as well as the type aB, keeps increasing towards .25, while the frequencies of the two original types decrease towards the same limiting ratio.

We can set up a spreadsheet model to examine this phenomenon. The reader who wishes to delve into an algebraic analysis of this model, as well as examine the case when the genes are on the same homologous pair (which can lead to the phenomenon of *linkage disequilibrium*) can consult exercises 16 and 17.

Example 12.9: Follow the three-step procedure outlined in the Hardy-Weinberg theorem to derive formulas for the new gene frequencies for a pair of genes at unlinked loci after one generation of random mating.

Solution:

We begin with a table giving the initial distribution of genotypes at the two loci:

Initial Gametic Frequency				
Gamete Type	AA	ab	aB	Ab
Frequency	p	q	s	t

We can now construct a table giving the frequencies of the nine possible zygotic genotypes (see Example 12.3). To do this we need to list for each zygote all possible parental genotype crosses that yield the given zygote genotype. When the parental types are distinct from each other there will be two possibilities for obtaining the cross by switching which type comes from the male and which from the female. This is indicated in the table below by writing $\times 2$ next to the appropriate cross.

Zygote Type	Parental Cross	Zygote Type	Parental Cross	Zygote Type	Parental Cross
$AA // BB$	$AB \times AB$	$Aa // BB$	$AB \times aB \times 2$	$aa // BB$	$aB \times aB$
$AA // Bb$	$AB \times Ab \times 2$	$Aa // Bb$	$AB \times ab \times 2$ $Ab \times aB \times 2$	$aa // Bb$	$aB \times ab \times 2$
$AA // bb$	$Ab \times Ab$	$Aa // bb$	$Ab \times ab \times 2$	$aa // bb$	$ab \times ab$

We can now express the zygotic genotypes in terms of the original gene frequencies. Using the previous table and the rules of probability we obtain the following table, which completes the 2nd stage of our analysis.

Zygote Type	Frequency	Zygote Type	Frequency	Zygote Type	Frequency
$AA // BB$	p^2	$Aa // BB$	$2ps$	$aa // BB$	s^2
$AA // Bb$	$2pt$	$Aa // Bb$	$2(pq + st)$	$aa // Bb$	$2qs$
$AA // bb$	t^2	$Aa // bb$	$2qt$	$aa // bb$	q^2

Table 12.1

(As a partial check of the analysis up to this point the reader might want to verify that the nine entries in the table have a sum of one. See exercise 15.) To complete the analysis we must find the new gamete frequencies from the distribution of zygotes. The table below lists the four gametes together with the zygotic types that can produce each one. The multiplicative factor following each entry indicates the fraction of the gametes produced by that zygote that would be of the stated type. This fraction is obtained using Mendel's law of independent assortment for each individual zygote.

Gamete Type	AB	ab	aB	Ab
From Zygote Types	$AA//BB\times 1$	$aa//bb\times 1$	$aa//BB\times 1$	$AA//bb\times 1$
	$AA//Bb\times \frac{1}{2}$	$aa//Bb\times \frac{1}{2}$	$aa//Bb\times \frac{1}{2}$	$AA//Bb\times \frac{1}{2}$
	$Aa//BB\times \frac{1}{2}$	$Aa//bb\times \frac{1}{2}$	$Aa//BB\times \frac{1}{2}$	$Aa//bb\times \frac{1}{2}$
	$Aa//Bb\times \frac{1}{4}$	$Aa//Bb\times \frac{1}{4}$	$Aa//Bb\times \frac{1}{4}$	$Aa//Bb\times \frac{1}{4}$

Using the zygote frequencies in Table 12.1, we can finally obtain the desired results for the new gene frequencies, which we denote by p', q', s' and t'. For example, the new frequency p' of AB is obtained by adding frequencies for the zygotes listed above under the column AB, each multiplied by the respective fraction. Thus

$$p' = p^2 + pt + ps + \tfrac{1}{2}pq + \tfrac{1}{2}st \tag{12.2}$$

The last equation can be put into a simpler form by using the fact that the four original frequencies satisfy $p+q+s+t=1$ or $p+t+s=1-q$. We can make use of this relationship in (12.2) by factoring p from the first three terms on the right side, obtaining

$$\begin{aligned} p' &= p(p+t+s) + \tfrac{1}{2}pq + \tfrac{1}{2}st \\ &= p(1-q) + \tfrac{1}{2}pq + \tfrac{1}{2}st \\ &= p - pq + \tfrac{1}{2}pq + \tfrac{1}{2}st \\ &= p + \tfrac{1}{2}(st - pq). \end{aligned}$$

In a similar manner we obtain the other formulas listed in the table below.

Gamete Type	AB	ab	aB	Ab
Frequency	$p' = p + \frac{1}{2}(st - pq)$	$q' = q + \frac{1}{2}(st - pq)$	$s' = s - \frac{1}{2}(st - pq)$	$t' = t - \frac{1}{2}(st - pq)$

Table 12.2

It is now a relatively trivial matter to setup a spreadsheet incorporating the above formulas. The next figure illustrates the appropriate entries. Row 1 contains text identifiers for the four genotypes. In row 2 you would type four specific numbers, not the generic letters shown. Of course these numbers must satisfy the condition that $p+q+s+t=1$. The formulas in row 3

encompass the results shown in Table 12.2. To study the genotype distribution through several generations, we have only to copy the formulas in row 3 into further rows below.

	A	B	C	D
1	*AB*	*ab*	*aB*	*Ab*
2	p	q	s	t
3	=A2+.5*(C2*D2-A2*B2)	=B2+.5*(C2*D2-A2*B2)	=C2-.5*(C2*D2-A2*B2)	=D2-.5*(C2*D2-A2*B2)

Let's apply this model to the situation mentioned earlier where $p = .5$, $q = .5$ and s and t are both initially zero. The table below gives the output of the model. As we indicated, all the gene frequencies approach a stable value, in this case .25. This limiting value can in fact be predicted from the gene frequencies at each locus given by the Hardy-Weinberg Theorem. See exercise 16 for details.

generation	AB	ab	aB	Ab
1	0.500	0.500	0.000	0.000
2	0.375	0.375	0.125	0.125
3	0.313	0.313	0.188	0.188
4	0.281	0.281	0.219	0.219
5	0.266	0.266	0.234	0.234
6	0.258	0.258	0.242	0.242
7	0.254	0.254	0.246	0.246

■

12.5 Summary

Because genetic recombination involves seemingly random selection, the tools of probability theory can be used to predict ensemble effects. Thus, while we cannot predict the outcome of an individual cross of two heterozygotes, we can accurately predict the probabilities for any particular outcome. In turn, these probabilities represent the relative frequencies with which individual outcomes will occur in a large number of repetitions of the reproductive process.

At the level of the individual, our analysis has focused on the zygotic distribution of genotypes at one or two loci. When the loci are on different chromosomes the analysis rests on Mendel's principle of independent assortment, according to which the selection of genes on different chromosomes constitute independent events. Additional mixing of genes occurs even on the same chromosome, due to the important phenomenon of **recombination**

Once we have worked out the probability distribution for zygotic genotypes, we can apply the results to determine the long-term distribution of genotypes in an entire population, provided we know the initial distribution. With respect to one locus, not subject to selection or assortive

mating, The **Hardy-Weinberg theorem** describes the distribution of genotypes. Similar, albeit more complicated patterns hold when we consider the interaction of two loci, whether on homologous or different chromosome pairs.

12.6 Exercises

1. The ABO blood type is a phenotype that is due to a gene with three alleles at a single locus. These are usually designated I^A, I^B, i where the allele I^A codes for the presence of antigen A, I^B codes for the presence of B and i provides no antigen. Each of the coding alleles is expressed regardless of any other allele with which it is paired. Based on your knowledge of blood types, determine which of the six genotypes correspond to each of the phenotype blood groups A, B, AB and O**.**

2. a) A father has type A blood and a mother type B. If one of their offspring has type O blood, what are the genotypes of the parents? (See exercise 1 regarding the genotype classification for the A, B, AB, O phenotypes.)

   b) A father with type A blood and a mother with type B have 4 children. The first three are of type AB and the last is of type A.

      i) What is the genotype of the mother?

      ii) What are the two possible genotypes for the father? For each genotype, find the probability of the specified pattern of blood types in the offspring. (Computations of this sort are used in the *maximum likelihood* method of statistical inference.)

3. a) How many genotypes are possible if there are 4 alleles at a locus?

   b) Give an argument to show that if there are n alleles then there are $n+\frac{n(n-1)}{2}$ genotypes.

4. Repeat the analyses of Example 12.3 and Example 12.4 if the two loci have three alleles at each locus.

5. Two genes each with two alleles, one of which is dominant, are on the same chromosome. If the two genes code for different traits, how many phenotypes are possible?

6. Suppose in Example 12.6 that A is dominant. Using the mating in that example, what are the probabilities that the zygote has each of the phenotypes A and a?

7. Five offspring are produced from male and female parents as described in Example 12.6. What is the probability that at least one of the five will be a heterozygote? What probabilistic assumption are you using in your calculation? (Remark: In Chapter 13 we will learn a technique (the binomial distribution) that enables you to find the probability for the occurrence of any number of offspring of a particular genotype.)

8. Assume in the situation described in Example 12.7 that A and B are each dominant. According to Example 12.5 there are four possible phenotypes. Find the probability of each phenotype for the mating in Example 12.7.

9. Assume as in Example 12.7 that the gene loci are on different chromosome pairs. Suppose the parents are of genotypes, male $Aa // BB$ and female $aa // Bb$. Find the probability for the genotype of all possible zygotes from such a pairing.

10. Suppose A and B are each dominant at their respective loci and each gene codes for some readily observable physical trait. The genotype of an individual cannot be inferred uniquely from the trait because AA and Aa individuals will exhibit the same character, namely phenotype A. If a male has phenotype AB what breeding experiment could you do to determine if the male was homozygous or heterozygous for each gene. In particular, what outcomes would distinguish the four possible genotypes for the male? Assume that you have access to females who are homozygous with respect to the recessive gene, i.e. of type $aa // bb$.

11. A certain species of plant has two genes A and B with two alleles (A, a) and (B, b) of which the A (respectively B) form is dominant. The genes may or may not be on the same homologous chromosome pair. We assume that the plant species is capable of self-fertilization from male and female organs. We start with a plant whose self-fertilized progeny all have phenotype AB and another plant whose self-fertilized progeny all have phenotype ab. These so-called pure breeding dominant and recessive types are crossed with each other. The offspring of the cross are known as the F_1 or first filial generation.

    Plants from the F_1 generation are then crossed with a pure breeding plant of phenotype ab. A particular experiment produced the following distribution of phenotypes.

Phenotype	AB	Ab	aB	ab
Frequency	140	75	65	120

    a) Explain why each of the original pure breeding parents must be homozygous at each locus and why the F_1 generation consists entirely of double heterozygotes.

    b) Show that the cross of the F_1 plants with the pure breeding recessive heterozygote produces zygotes with only four distinct genotypes, regardless of whether the genes A and B are on the same chromosome pair or different pairs. In each case one genotype corresponds exactly to one phenotype.

    c) Why is the phenotype data above inconsistent with the hypothesis that the genes A and B are on different chromosome pairs?

    d) Which phenotype combinations correspond to crossovers? Estimate the recombination frequency from the given data.

12. Sickle-cell trait is a defect in the shape of the hemoglobin molecule, which leads to an often fatal anemia. It is caused by a recessive allele S of a polymorphic gene. The other alleles

code for a normal hemoglobin shape, are all dominant over the allele S and for the purpose of this analysis can be thought of as a single alternative which we denote by A. A study of 12,387 adults from the Yoruba people of Nigeria yielded the following data on the three genotypes SS, SA, and AA. (Source: *Evolution*, 2nd ed., by M. Ridley)

Genotype	SS	SA	AA
Frequency	29	2993	9365

Show that the population is not in Hardy-Weinberg equilibrium with respect to this locus. It is known in this instance that the heterozygote has a survival advantage over either homozygote. The genotype frequencies will approach an equilibrium value, but not the one predicted by the Hardy-Weinberg theorem.

13. Suppose that there are three alleles A_1, A_2, A_3 at the locus and that conditions (i) to (v) listed for the Hardy-Weinberg theorem (Theorem 12.2) are in effect. Denote the frequencies of these genes in the gene pool by $p, q,$ and r, respectively.

    a) Compute the frequencies of the six genotypes. (See Example 12.2)

    b) Using the six genotype frequencies found in a) find the three gene frequencies in the next adult generation.

    c) What are the stable genotype frequencies (Hardy-Weinberg ratios) for the case of three alleles?

14. A single homologous pair of chromosomes determines the sex of a human zygote. Males carry a single X chromosome paired with a slightly shorter Y chromosome. Females carry a pair of X chromosomes. A gene that appears on the X chromosome, but not on the shorter Y is said to be sex-linked. Males carry only a single allele of such a gene. In this exercise we show how the population gene frequencies of a sex-linked gene change with random mating and explore a numerical simulation of this process. We will consider the mathematics further in chapter 17.

    a) We begin with the first stage of the analytical scheme in Figure 12.1, leaving some details to the reader. Suppose a gene with two alleles A and a is sex-linked. We need to consider two separate gene pools, one for the males and one for the females.

	Initial Male Gene Pool		**Initial Female Gene Pool**	
Gamete	A	a	A	a
Frequency	p	q	s	t

    Why is $p+q=\frac{1}{2}$ but $s+t=1$?

    b) The next step is to deduce the zygote frequency table for male zygotes and female zygotes. Derive the entries shown in the following table.

Male Zygotes			Female Zygotes		
genotype	*AY*	*aY*	*AA*	*Aa*	*aa*
frequency	*s*	*t*	$2ps$	$2(pt+qs)$	$2qt$

Hint: Each of the stated frequencies must be computed as a conditional probability. Thus the frequency (probability) entry for male zygotes of type *AY* refers to the fraction of <u>males</u> that are of this type. In probability terms this is $P(AY \mid \text{male})$.

c) Using a) verify that the zygote frequencies for each sex listed above have a sum of one.

d) We now derive the new gametic frequencies. Explain why these are given by the following table:

	New Male Gene Pool		New Female Gene Pool	
Gamete	*A*	a	*A*	*a*
Frequency	$p' = \frac{1}{2}s$	$q' = \frac{1}{2}t$	$s' = p + \frac{1}{2}s$	$t' = q + \frac{1}{2}t$

e) The file *hw_sex.xls* provides a framework for examining the frequency changes of a sex-linked gene. Follow the steps given in that spreadsheet and answer the questions regarding the limiting values of the gene frequencies.

15. Verify the assertion made earlier that the sum of the nine frequency entries in Table 12.1 is one.

16. a) Set up the spreadsheet model for the changes in gene frequencies on non-homologous chromosomes as described in Example 12.9. Determine the limiting gene frequencies in each of the following cases:

i) $p = 0, \quad q = 0.1, \quad s = 0.1, \quad t = 0.8$

ii) $p = 0.1, \quad q = 0, \quad s = 0.1, \quad t = 0.8$

iii) $p = 0.1, \quad q = 0.1, \quad s = 0, \quad t = 0.8$

b) Explain why the gene frequencies of *A*, *a*, *B*, and *b* are given respectively by $p_A = p + t$, $p_a = q + s$, $p_B = p + s$, and $p_b = q + t$. By the Hardy-Weinberg Theorem (Theorem 12.2) these frequencies should remain fixed from one generation to the next. Verify this using the formulas in Table 12.2.

c) For the numerical examples in a) verify that the limiting value of p is $p_A p_B$, the limiting value of q is $p_a p_b$, the limiting value of s is $p_a p_B$, and the limiting value of t is $p_A p_b$. In other words, after several generations the population gene frequencies at two loci can be found by multiplying the frequencies at each locus. This is a fundamental principle used in computing frequencies of DNA matches in forensic DNA analysis.

d) Using that $s = p_B - p$ and $t = p_A - p$ (see b)), show that $st - pq = p_A p_B - p$. From Table 12.2 conclude that $p' - p_A p_B = \frac{1}{2}(p - p_A p_B)$ and that after the next generation $p'' - p_A p_B = \frac{1}{4}(p - p_A p_B)$. By continuing this argument, explain why the frequency of the AB gamete will approach $p_A p_B$. Derive similar results for the other gamete frequencies.

17. The analysis of population gene frequencies for genes on the same chromosome pair is similar to the analysis given in Example 12.9 and exercise 16. Assume that A, a, B, and b are on the same homologous pair, with a recombination frequency r (see Theorem 12.1). We let p, q, s, and t denote the frequencies of the gametes AB, ab, aB, and Ab, respectively.

a) Construct a table similar to Table 12.1 showing the frequency of all (ten) possible zygotes.

b) Using the table found in a) construct a table similar to Table 12.2 for the new gametic frequencies. Your formulas should be identical to those in Table 12.2, except that the factor $\frac{1}{2}$ will be replaced by the recombination frequency, r.

c) Set up a spreadsheet model for the changes in gene frequency as described in b). Use absolute referencing to refer to the value of *r*. Explore the situation where the initial gene frequencies are $p = q = 0.5$, $s = t = 0$ and where *r* takes on the values $.01, .05, .1, .25$ and $.5$. Verify that for each value of *r* the limiting gene frequencies are the same (and equal to the product of the frequencies at each locus) but that the number of generations required to "reach" the limit increases as the recombination frequency gets smaller. The long time required to reach "equilibrium" when the loci are very close leads to what biologists refer to as *linkage disequilibrium*.