6

Genetics

We have all observed that offspring tend to have physical traits in common with their parents. In humans, similarity in hair color, eye color, height, and build often quite clearly run in families. That selective breeding might enhance traits must have been noticed long ago in our history, as domesticated animals and crops have strongly developed features that we find useful.

On the other hand, the traits of offspring are generally not completely predictable from observing those of the parents. A child might have a trait, such as hemophilia, that neither parent exhibits, though such a trait might occur more commonly within one family than another. Thus, despite patterns to inheritance, chance also appears to be involved. Creating a mathematical model of heredity requires capturing both of these aspects.

The first decisive step was taken by the Augustinian monk Gregor Mendel in the latter half of the nineteenth century. Experimenting with some carefully chosen traits in peas, he was led to propose what we now call a *gene* as the basic unit of inheritance. Though it is perhaps surprising to the modern student, at that time the gene was an entirely abstract concept, with no proposed physical basis, such as the DNA sequences we now immediately imagine.

Recognizing the value of quantitative analysis, Mendel created a mathematical model for the transmission of heritable traits, based on the concepts of probability. His genius was in both identifying simple enough traits to be able to formulate a good model and then modeling the inheritance of those traits successfully. Though subsequent work has added many new features to our models, and we now know much more about the chemical and biological mechanisms behind genetics, Mendel's simple model remains the basic core of our understanding of how many organisms pass on traits to their offspring.

6.1. Mendelian Genetics

In 1865, Mendel presented his findings from breeding experiments with garden peas (Mendel, 1866) to a small group of scientists in Brünn, in the

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

modern-day Czech Republic. Although the world scientific community largely failed to notice until the turn of the century, Mendel's genetic theory was a major advance. Let's consider some of his experiments carefully to understand how the model describes what he observed.

Mendel isolated seven characteristics of pea plants: stem length, seed shape, seed color, flower color, pod shape, pod color, and flower position for study. Each of these characteristics appeared in the peas in one of two forms we'll call *traits*. For instance, stem length might be tall or dwarf, while seed shape could be round or wrinkled. By selective breeding, he then developed *true-breeding* lines of peas for these traits – strains of pea plants that produced progeny, all of which were identical to the parents. Thus, all the descendents of a true-breeding line for tall plants would be tall, and all the descendents of a true-breeding dwarf line would be dwarf.

For each of the characteristics, Mendel cross-bred the two true-breeding lines. For example, true-breeding tall plants were crossed with true-breeding dwarf plants, and true-breeding plants with smooth seeds were crossed with true-breeding plants with wrinkled seeds. Thus, inheritance could be studied one characteristic at a time, and the influence of pure parental traits on the progeny observed. Mendel discovered that, in these crosses, the progeny displayed only one of the traits of the parental generation: The progeny of tall and dwarf plants were all tall; the progeny of plants with round seeds and those with wrinkled seeds all had round seeds. Since the same trait from the parental generation was exhibited by all the progeny, Mendel called such a trait *dominant* and the hidden trait *recessive*. The dominant traits discovered by Mendel's crosses are given in Table 6.1.

Mendel furthered experimented by allowing the offspring of these first generation crosses, or F_1, to self-pollinate and produce a second generation F_2. (The symbols F_1 and F_2 are the standard notations in genetics for the first and second filial generations.) Interestingly, the recessive traits, absent in F_1,

Table 6.1. *Mendel's F_1 Data*

Parental Traits	Dominant Trait
Tall, dwarf plants	Tall
Round, wrinkled seeds	Round
Yellow, green seeds	Yellow
Purple, white flowers	Purple
Inflated, constricted pods	Inflated
Green, yellow pods	Green
Axial, terminal flowers	Axial

Table 6.2. *Mendel's F_2 Data*

Cross Producing F_1	F_2	Ratio
Tall × dwarf plants	787 tall, 277 dwarf	2.84:1
Round × wrinkled seeds	5,474 round, 1,850 wrinkled	2.96:1
Yellow × green seeds	6,022 yellow, 2,001 green	3.01:1
Purple × white flowers	705 purple, 224 white	3.15:1
Inflated × constricted pods	882 inflated, 299 constricted	2.95:1
Green × yellow pods	428 green, 152 yellow	2.82:1
Axial × terminal flowers	651 axial, 207 terminal	3.14:1

reappeared in F_2. Mendel's data for the frequency of each observed trait is shown in Table 6.2.

The last column of Table 6.2 shows the ratio *(number of plants with dominant trait):(number of plants with recessive trait)* in the F_2 plants. These ratios are all remarkably close to 3:1 for each of the seven traits under study. (In fact, they are so close to 3:1 that some believe Mendel may have selectively reported his data at a time when scientific standards were less developed.)

► Is noticing this 3:1 ratio enough to help you create an entire genetic theory, as Mendel did?

To explain the 3:1 ratio, Mendel proposed that, for each characteristic, a pea plant must contain a pair of the hereditary factors now called genes. Each gene can come in several forms or *alleles*, corresponding to variations within a trait. For example, for the stem length trait, there is a dwarf allele, d, and a tall allele, D. (Usually, we choose a small letter for a gene based on the recessive allele and use the corresponding capital letter for the dominant allele.) The true-breeding strains of pea plants contain two identical alleles and are said to be *homozygous*. The *genotypes* of these strains are dd for the dwarf strain and DD for the tall strain.

Mendel hypothesized that each parent passed along exactly one of its genes to its progeny. If a parent has genotype Dd, either a tall D allele or a dwarf d allele is passed on, rather than some sort of mix of the two. This *principle of segregation* treats the alleles associated with traits as discrete and indivisible units. A further consequence of the principle is that progeny will also have exactly two genes for a characteristic, as did their parents, and thus the number of genes does not increase in successive generations.

Chance is introduced into the model in determining which of the parental genes each descendent receives. With equal probability, either of the genes in the father will be passed to a descendent, and with equal probability, either of

the genes in the mother will be passed on as well. It's as if two parental coin flips determine the outcome in the progeny.

Much more is now known about how genes segregate in the formation of *gametes* or reproductive cells. *Meiosis* is a complicated process in which gametes (egg and sperm in animals, spores in plants) carrying only one copy of each gene are formed. Modern understanding is that genes are found arranged linearly on *chromosomes*, large molecules residing in the nucleus of cells. The chromosomes come in pairs, accounting for the two copies of each gene. Indeed, in gamete formation, it is chromosomes that segregate, not genes as Mendel proposed. At fertilization, two gametes, each carrying one copy of each chromosome, join to produce new offspring.

In reality, inheritance of chromosomes is much more complicated than can be captured by our Mendelian model. The process of *crossing over*, an important source of genetic variability, makes segregation quite involved. Moreover, not all alleles fit the dominant/recessive framework that the Mendelian model supposes, and many traits are not determined by a single gene, but rather by collections of genes. Finally, whereas most familiar organisms do carry two copies of each gene in most cells, and are thus called *diploid*, there are exceptions to this.

However, we are getting ahead of ourselves by bringing up all these complications. The Mendelian model is remarkably good for predicting and understanding the inheritance of many traits and marks a first step toward understanding the biology of inheritance. We can bring modifications into the model later, after we fully understand Mendel's simpler view. So, for now, we will continue to assume segregation of parental genes and restrict our attention to the situation in which a single gene controls a single trait.

When Mendel crossed the DD true-breeding tall genotype with the dd true-breeding dwarf genotype, each descendent inherited an identical set of genes from the parents: D from the first parent and d from the second. Genotypically, these progeny are all Dd, and, because they contain two different forms of the gene, are said to be *heterozygous*.

Remember that each of the F_1 were tall pea plants. Thus, although genetically the progeny were heterozygous, the D allele was dominant over the d allele, in the sense that all the plants of F_1 resembled their tall parent. These F_1 have the same *phenotype* as their tall parents, that is, they have the same observable characteristics.

► What are the phenotypes of the genotypes DD, Dd, and dd?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.3. *Punnett Square for* $Dd \times Dd$

	D	d
D	DD	Dd
d	Dd	dd

► If W denotes the dominant allele for round seeds and w the recessive allele for wrinkled seeds, what are the phenotypes of WW, Ww, and ww?

To understand the 3:1 phenotypic ratio in F_2, a helpful device is the *Punnett square*. Here, we place the possible gametes formed by F_1 parents as row and column headings. The entries, formed by the union of such gametes, are the F_2 genotypes. A Punnett square for the stem length gene in the self-fertilization of a pea in F_1 is shown in Table 6.3.

Because each of the gametes is equally likely, according to the model, the four entries of the square are all equally likely descriptions of the genotypes of offspring. We should thus find the three genotypes DD, Dd, and dd in a ratio of 1:2:1 in the F_2 plants.

Notice that we can also deduce the expected ratio of the phenotypes of the F_2 progeny. Since the first two of these genotypes produce the tall phenotype, we should see three tall plants (DD and Dd) for every dwarf plant (dd), giving a ratio of 3:1. Mendel's simple genetic model describes the outcome of his breeding experiments remarkably well.

We can easily extend Mendel's model to make predictions about the outcome of more complicated breeding experiments. For example, if W and w denote the alleles for round and wrinkled seeds, then we may be interested in predicting the outcome of the cross $DdWw \times ddWw$. To handle such two gene crosses, we assume, as Mendel did, that genes *assort independently*. That is, in gamete formation, the segregation of alleles of one parental gene occurs independently of the segregation of alleles for the other gene. Using the language of probability, we would say the segregations of the alleles for two different genes are independent events.

Example. To predict the outcome of the cross $DdWw \times ddWw$, we can again use a Punnett square. Because all combinations are equally likely in gametes, the parental type $DdWw$ creates four types of gametes – DW, Dw,

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.4. *Punnett Square for* $DdWw \times ddWw$

	DW	Dw	dW	dw
dW	$DdWW$	$DdWw$	$ddWW$	$ddWw$
dw	$DdWw$	$Ddww$	$ddWw$	$ddww$
dW	$DdWW$	$DdWw$	$ddWW$	$ddWw$
dw	$DdWw$	$Ddww$	$ddWw$	$ddww$

dW, and dw – all with equal probability. Similarly, $ddWw$ creates gametes dW, dw, dW, and dw with equal probability. The resulting Punnett square is shown in Table 6.4.

Notice that there are only six different genotypes among the 16 entries in the square. However, since several of these genotypes produce the same phenotype, there are only four different phenotypes represented. Careful counting shows that, among the F_2 plants, we should expect the following fractions of the population with the given phenotypes: Tall plants with round seeds $6/16$, tall plants with wrinkled seeds $2/16$, dwarf plants with round seeds $6/16$, and dwarf plants with wrinkled seeds $2/16$.

- Which genotypes in the square produce the phenotype of tall plants with wrinkled seeds?
- Suppose you wanted to determine the phenotypes and their frequencies for a cross between $DdWwYY \times ddWWYy$, where Y represents the dominant allele for green pods and y the recessive allele for yellow pods. How big would your Punnett square be?

The size of a Punnett square grows quickly, in fact exponentially, with the number of independently assorting genes you are tracking. For n genes, the square is $2^n \times 2^n$. This makes Punnett squares impractical for all but the simplest of analyses. Ultimately, we will find that the language of probability, as introduced in Chapter 4, is a better tool for calculating the chances of different outcomes of a particular cross.

Example. To redo the example above of the cross $DdWw \times ddWw$ using probability, let's first calculate the likelihood of dwarf progeny with wrinkled seeds. Because both dwarfness and wrinkled seeds are recessive characteristics, we know that the only genotype producing these traits is $ddww$. This means that each parental strain must contribute a d and a w to such progeny.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

In detail,

$$\begin{aligned}
&\mathcal{P}(\text{dwarf plants with wrinkled seeds}) \\
&\quad = \mathcal{P}(ddww) \\
&\quad = \mathcal{P}(dw \text{ from first parent})\mathcal{P}(dw \text{ from second parent}) \\
&\quad = \mathcal{P}(d \text{ from first parent})\mathcal{P}(w \text{ from first parent}) \\
&\qquad \times \ \mathcal{P}(d \text{ from second parent})\mathcal{P}(w \text{ from second parent}) \\
&\quad = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(1\right)\left(\frac{1}{2}\right) = \left(\frac{1}{8}\right).
\end{aligned}$$

Naturally, this answer agrees with our result from the Punnett square.

Let's pause and examine several of the equal signs above, since they are derived from important mathematical and biological concepts. Note, for example, that the second equality, rewriting $\mathcal{P}(ddww)$ as the product of the probabilities of inheriting alleles from each parent, is only correct if these are independent events. This assumption of independence is part of the notion of *random union of gametes*: The probability of any union of paternal and maternal gametes is the product of the proportions in which those gametes occur.

- Why biologically should what is inherited from one parent be independent of what is inherited from the other?

In addition, the third equality, writing the probability of inheriting dw from a parent as the product of the probabilities of inheriting d and w, is a mathematical restatement of the principle of independent assortment.

Let's try another, more involved, example.

Example. What is the probability that the progeny of $DdWw \times ddWw$ is dwarf with round seeds?

Because having round seeds is a dominant characteristic, two genotypes WW and Ww, both give rise to round seeds, and we will need to take both possibilities into consideration. Now

$$\begin{aligned}
\mathcal{P}(\text{dwarf with round seeds}) &= \mathcal{P}(ddWW \text{ or } ddWw) \\
&= \mathcal{P}(ddWW) + \mathcal{P}(ddWw),
\end{aligned}$$

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

since $ddWW$ and $ddWw$ are disjoint events. But

$$\begin{aligned}\mathcal{P}(ddWW) &= \mathcal{P}(dW \text{ from first parent})\mathcal{P}(dW \text{ from second parent}) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(1)\left(\frac{1}{2}\right) = \left(\frac{1}{8}\right),\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}(ddWw) &= \mathcal{P}(dW \text{ from first parent})\mathcal{P}(dw \text{ from second parent}) \\ &\quad + \mathcal{P}(dw \text{ from first parent})\mathcal{P}(dW \text{ from second parent}) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(1)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(1)\left(\frac{1}{2}\right) \\ &= \left(\frac{1}{8}\right) + \left(\frac{1}{8}\right) = \left(\frac{1}{4}\right).\end{aligned}$$

Finally, we add to compute

$$\mathcal{P}(\text{dwarf with round seeds}) = \left(\frac{1}{8}\right) + \left(\frac{1}{4}\right) = \left(\frac{3}{8}\right).$$

► Justify each step of these computations. Where have we used the fact that certain events are independent?

The last calculation was complicated enough that it is a good idea to check it a different way. Let's do this by thinking of the principal of independent assortment differently, in more probabilistic terms. When we say that the two genes assort independently, we mean that the events $E_d = \{\text{plant is dwarf}\}$ and $E_r = \{\text{plant has round seeds}\}$ are independent events. Thus, we can compute probabilities for each of these events, and then use the multiplication rule for independent probabilities to combine the answers. This focuses our attention on more manageable problems; instead of looking at the cross $DdWw \times ddWw$, we can look at the crosses $Dd \times dd$ and $Ww \times Ww$ separately.

Example. To find the probability of dwarf progeny with round seeds, we need only multiply the probabilities:

$$\mathcal{P}(\text{dwarf with round seeds}) = \mathcal{P}(E_d \cap E_r) = \mathcal{P}(E_d)\mathcal{P}(E_r).$$

For the $Dd \times dd$ cross, $\mathcal{P}(E_d) = \mathcal{P}(dd)$. This probability is $\mathcal{P}(dd) = \frac{1}{2}$, which could be found either by using a 2×2 Punnett square or by arguing

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

that

$$\mathcal{P}(dd) = \mathcal{P}(d \text{ from first parent})\mathcal{P}(d \text{ from second parent}) = \left(\frac{1}{2}\right)(1) = \frac{1}{2}.$$

For the $Ww \times Ww$ cross,

$$\mathcal{P}(E_r) = \mathcal{P}(WW \text{ or } Ww) = \mathcal{P}(\text{not } ww) = 1 - \mathcal{P}(ww) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Thus,

$$\mathcal{P}(E_d \cap E_r) = \left(\frac{1}{2}\right)\left(\frac{3}{4}\right) = \frac{3}{8},$$

just as we found before.

While we have seen several ways to calculate the frequencies of phenotypes in progeny from various crosses, we will use probability most often. It is a sophisticated tool that allows us to estimate and model genotypic and, more generally, allelic frequencies in a population. As we move into increasingly complicated genetic models, simple devices like the Punnett square cannot be usefully adapted.

Although the basic Mendelian model does not describe all genetic phenomena of interest, it is adequate to model the incidence of certain human diseases. For example, Tay-Sachs disease, a disease primarily striking children of Ashkenazi Jewish descent and usually leading to death before age 5, is developed by individuals who are homozygous with a recessive allele for a particular gene. It is estimated that roughly 1 in 31 adults in the Ashkenazi population in North America are heterozygous for the recessive allele. Recessive alleles may lie hidden for generations if most individuals marrying into a family are homozygous dominant. Tay-Sachs disease may thus occur unexpectedly and with devastating impact when two heterozygous individuals have children. For many years, estimates of the presence of the recessive allele in the population, together with extensive family medical histories when they were available, were the only means of calculating the risk that a child would develop Tay-Sachs disease. More recently, prenatal techniques such as amniocentesis are used to detect the presence of the Tay-Sachs mutation.

Problems

6.1.1. Imagine that, in a certain species, gamete formation does not occur, and instead of receiving half the genes of each parent, offspring receive the *full* set of genes from both parents. If each parent in the

founding generation F_0 has two copies of a particular gene, how many copies will offspring of the nth generation F_n have?

6.1.2. Create a Punnett square for a $DdWw \times DdWw$ cross of pea plants. What proportion of the progeny has each genotype? Each phenotype?

6.1.3. In the text, probabilistic arguments are given to compute the probability of dwarf wrinkled-seed and dwarf round-seed phenotypes for the progeny of a $DdWw \times ddWw$ cross of pea plants. Complete the analysis by using probability to compute the following:
   a. The probability of a tall wrinkled-seed phenotype.
   b. The probability of a tall round-seed phenotype.

6.1.4. According to (Petersen *et al.*, 1983), the recessive allele for Tay-Sachs disease is present in 1 of 31 people in the North American Jewish subpopulation. Because of the nature of the disease, we can assume all adults with the allele are heterozygous.
   a. What is the probability that a couple drawn from this subpopulation will both have the allele?
   b. What is the probability that a child of such a couple will develop Tay-Sachs disease?
   c. What is the probability that a child, both of whose parents come from this subpopulation, will develop Tay-Sachs disease?

6.1.5. Consider three genes, each with dominant and recessive alleles, denoted A, a; B, b; and C, c.
   a. If an individual has genotype $AaBbCC$, how many different gametes might it form?
   b. If two organisms with genotype $AaBbCC$ are mated, how many different genotypes and phenotypes are possible?

6.1.6. Generalize the result of the last problem by considering N genes for a particular organism, with each gene having dominant and recessive alleles.
   a. How many different gametes can be formed by an organism that is heterozygous for n genes and homozygous for $N - n$ genes?
   b. Suppose two individuals, with identical genotypes, are heterozygous for n genes and homozygous for $N - n$ genes. How many different genotypes and phenotypes are possible if these two organisms of identical genotype are mated?
   c. Suppose one individual is heterozygous for the first k genes only and a second individual is heterozygous for first l genes only, where $k < l \leq N$. At the other loci, both organisms are homozygous recessive. How many genotypes are possible when these organisms are crossed? How many phenotypes?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

6.1.7. A *testcross* is a cross between a genetically unknown organism and a homozygous recessive organism. Testcrosses can be used to determine whether an organism is heterozygous or homozygous for a particular allele.

a. Suppose three organisms with genotypes AA, Aa, and aa are crossed with aa. What is the expected ratio of phenotypes in each of these testcrosses?

b. Suppose that a pea plant of an unknown genotype is testcrossed with dwarf pea plants that have wrinkled seeds and yellow pods ($ddwwyy$). Of the progeny, some are tall, some are dwarf, and all have wrinkled seeds with green pods. What is the genotype of the unknown parental strain?

c. Explain why you can determine the genotype of an unknown plant by testcrossing with another plant that is homozygous recessive for all genes of interest, but not by testcrossing with a plant that is homozygous dominant. Give both informal reasoning and quantitative justification.

6.1.8. In rabbits, two independently assorting genes affect fur. The dominant allele, B, determines black fur, and a recessive allele b determines brown fur. Normal fur length is determined by a dominant allele, R, and short fur length by a recessive allele, r. A homozygous (in both genes) black rabbit with normal-length fur is crossed with a brown, short-haired rabbit.

a. What are the possible genotypes and phenotypes of the offspring in F_1? What is the proportion of each?

b. If F_1 rabbits are intercrossed, what proportion of the F_2 are homozygous (both dominant and recessive) for the color gene? What proportion are homozygous for both genes? What proportion of the black rabbits are homozygous for both genes?

c. What is the genotype ratio for black rabbits with normal length fur in F_2?

6.1.9. To test his hypothesis that two genes assorted independently, Mendel carried out another series of experiments. In one, he bred true-lines of pea plants with round, yellow seeds ($WWGG$) and wrinkled, green seeds ($wwgg$). Here, the recessive allele for green seed color is denoted by g. He crossbred these lines to get F_1, and then the F_1 plants self-fertilized to produce F_2.

a. What are the phenotypes and genotypes of F_1?

b. What phenotypes will be represented in F_2, and in what relative frequencies should the phenotypes occur if the genes do assort independently?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

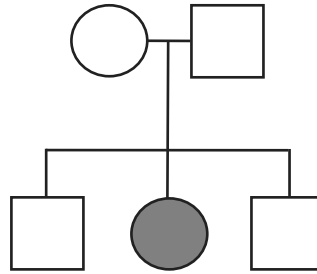

Figure 6.1. A human pedigree.

c. If Mendel's data did not exactly match that predicted in part (b), should he have doubted the assumption of independent assortment? By how much could the frequency data be off from the theoretical prediction before you would doubt the assumption? Explain informally.

6.1.10. If a certain genotype is lethal to embryos, then the expected proportions of genotypes in a new generation is, of course, affected.

In mice, an allele Y^l, known as the *yellow-lethal* mutation, is dominant for yellow fur color, but homozygotes die in the embryonic stage. Homozygotes with genotype yy have gray-brown or agouti fur.

Suppose two yellow mice are crossed. Give the genotypes, phenotypes, and expected proportions of their viable progeny, F_1.

6.1.11. Family pedigrees can be used in determining the risk of human offspring developing certain genetic diseases. One such disease is sickle-cell anemia, which occurs in individuals homozygous for a certain recessive allele.

In the pedigree of Figure 6.1, circles denote females and squares males; horizontal lines join couples, and vertical lines indicate children. Gray coloring indicates an individual has sickle-cell anemia.

a. For the relevant gene, what must the genotypes of the parents be?
b. What is the probability that a fourth child of the parents will be disease-free?
c. What are the possible genotypes of one of the sons? What is the probability of each of those genotypes?

6.1.12. Brachydactyly, or short fingers, is determined in humans by a particular gene with dominant and recessive alleles. Suppose a couple, both with brachydactyly, have two children. One child has normal length fingers and the other has short fingers.

a. Is brachydactyly a dominant or recessive trait? What are the genotypes of the parents? Of the children?
b. Suppose the couple has two more children. What is the probability that neither of them will have brachydactyly?

6.1.13. Plants heterozygous for three independently assorting genes are crossed.
a. What proportion of the progeny is expected to be homozygous for all three dominant alleles?
b. What proportion of the progeny is expected to be homozygous for all three genes?
c. What proportion of the progeny is expected to be homozygous for exactly one gene?
d. What proportion of the progeny is expected to be homozygous for at least one gene?

6.1.14. Mendel's simple model of dominant and recessive alleles does not always apply. Even when one gene with two alleles controls a trait, sometimes neither is completely dominant.

For example, in snapdragons, homozygous WW have red flowers and ww have white flowers. In heterozygotes Ww, however, both genes are expressed, and the flowers are pink. In such a case, the alleles are said to be *partially dominant*. If the heterozygote's phenotype is midway between those of the homozygotes, we say the alleles are *semidominant*.

For snapdragons, what are the phenotypic proportions in F_1 resulting from a $WW \times ww$ cross? What are the phenotypic proportions in F_2 arising from F_1 self-fertilization?

6.1.15. Some genes have multiple alleles, that is, more than two alleles exist in a population for a gene at a particular locus.

Suppose a gene has alleles a_1, a_2, and a_3, and that a_1 is dominant over a_2 and a_3, and a_2 is dominant over a_3. What are the genotype and phenotype frequencies you would expect from a cross $a_1a_3 \times a_2a_3$?

6.1.16. Mendel's model may be modified as in the last two problems to account for a gene that has more than two alleles, some of which exhibit partial or semi-dominance.

For example, the three alleles for human blood type – I^A, I^B, and I^0 – exhibit both dominance and *codominance*. Both I^A and I^B are dominant over I^0, but an individual with genotype I^AI^B will have type AB blood, because both alleles are expressed equally.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

a. What are the possible genotypes for the four phenotypic blood types, A, B, AB, and O?
b. Suppose an individual homozygous with type A blood marries an individual heterozygous with type B blood. What are the possible phenotypes of any offspring, and in what relative frequencies do these occur?
c. Suppose parents, heterozygous with types A and B blood, have four children. How many of the children would you expect to have type O blood? Would it be possible for all of the couple's children to have type O blood? Explain, both informally and quantitatively.

6.1.17. Mendel's basic model only describes phenotypic traits that are controlled by a single gene. However, most traits are more complicated.

A classic example is comb shape in chickens, which is determined by two independently assorting genes. There are four shapes of chicken combs: rose, pea, single, and walnut. Two genes with two alleles each are responsible for comb shape. The genotypes of the four shapes are: rose $R\text{–}pp$, pea $rrP\text{–}$, single $rrpp$, and walnut $R\text{–}P\text{–}$. (Here, a dash indicates either a dominant or a recessive allele is possible.)

a. What phenotypes result from the crosses $RRpp \times rrpp, rrPP \times rrpp$?
b. What phenotypes, and in what proportions, result from a $RRpp \times rrPP$ cross? If the F_1 progeny are interbred, what phenotypes, and in what proportions, are represented in F_2?

6.2. Probability Distributions in Genetics

While the Mendelian model gives a good understanding of the probability of a single child of certain parents being homozygous recessive for a particular gene, or of a single F_2 plant being tall or dwarf, often we are interested in calculating probabilities of more complicated events. For some of these, we need additional knowledge of probability, rather than genetics.

The term "*random variable*" is sometimes used for the outcome of a measurement or count when we believe some sort of random process underlies the experiment. A few examples of random variables are

- A fair coin is flipped 10 times, and the number of heads is counted. This number is a random variable that might take on the values $0, 1, 2, \ldots, 10$.
- Parents who are heterozygous for the Tay-Sachs recessive allele have three children. The number of their children that are homozygous recessive is a random variable that might take on the values 0, 1, 2, or 3.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

- Mendel's F_2 data on the progeny of the self-fertilized F_1 cross between tall and dwarf pea plants involved 1,064 plants. The proportion of the plants in F_2 that are tall was a random variable that could have taken on any of the values $0 = 0/1064$ (if all plants were dwarf), 1/1064, 2/1064, ..., $1 = 1064/1064$ (if all plants were tall).

The random variables listed here are built by counting or finding proportions of outcomes of simpler events. Our understanding of the simpler events should enable us to analyze these, but how?

A function that describes the probability of the various outcomes of a random variable is called a *probability distribution*. In this section, we will consider two particular distributions of use in genetics.

The binomial distribution and expected values. As a first example, suppose we flip a fair coin 3 times, and are interested in the probability of getting exactly 2 heads among the 3 flips. We can list the 8 equally likely outcomes of the 3 coin flips,

$$HHH, HHT, HTH, HTT, THH, THT, TTH, TTT,$$

each occurring with probability 1/8. In this list, the 3 outcomes

$$HHT, HTH, THH$$

have exactly 2 heads. Thus, using the addition rule of probabilities, we find

$$\mathcal{P}(\text{exactly 2 heads in 3 flips}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 3\left(\frac{1}{8}\right) = \frac{3}{8}.$$

Now suppose we wanted to find the probability of exactly 12 heads in 35 flips? We could proceed similarly, but listing cases is likely to be difficult and error-prone. However, a probability distribution called the *binomial distribution* allows us to calculate such probabilities quickly and efficiently, by associating probabilities to each of the possible outcomes $0, 1, 2, \ldots, n$ of the random variable that gives the number of heads produced by n coin flips.

To develop a formula for the binomial distribution, let's examine the example of 2 heads in 3 flips again. For each coin flip, we have probability 1/2 of getting a head and probability 1/2 of getting a tail. Thus, for any particular way we might get 2 heads and a tail, the probability is

$$\mathcal{P}(HHT) = \mathcal{P}(HTH) = \mathcal{P}(THH) = \left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{1}{8}.$$

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

The 2 heads required two factors of $1/2$, and the single tail required an additional factor of $1/2$. Because the coin flips are independent, we multiply these factors.

Now why are there three scenarios in which 2 heads can be produced in the 3 flips? We have to account for which of the 3 particular flips were the ones in which the heads occur. They could occur on flips 1 and 2, flips 1 and 3, or flips 2 and 3. The number of scenarios is the number of different ways that 2 of the 3 flips can be designated as producing heads.

This simple example motivates the general formula for the binomial distribution. Suppose we perform n independent trials of a random process that has two possible outcomes. For convenience, we call one of the two outcomes a *success* S and the other a *failure* F. Suppose further that in each trial we have $\mathcal{P}(S) = p$ and $\mathcal{P}(F) = q = 1 - p$. Then, the binomial distribution calculates the probability of k successes among the n trials, as

$$\mathcal{P}(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k q^{n-k}.$$

Here, we have introduced the notation $\binom{n}{k}$ to mean the number of different ways that the k successes might be located among the n trials.

- Above, we calculated the probability of 2 heads in 3 coin flips. What is a success? A failure? What is the number of trials n and the number of successes k?

Of course, for the binomial formula to be useful, we need a good way to find a value for $\binom{n}{k}$. For the 3-coin flip example, thinking of "heads" as a success, we computed $\binom{3}{2} = 3$ by listing all the cases. (Alternatively, if we think of "tails" as a success, the same list shows $\binom{3}{1} = 3$.) It really is not feasible to list all the possibilities for large n and k, however. In the exercises, you will develop the formula

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}. \tag{6.1}$$

The expression $\binom{n}{k}$ is called the number of *combinations* of n objects chosen k at a time, but is usually read as "*n choose k*." We think of it as counting how many ways we can designate (or choose) k out of the n trials to be the ones where the successes occur.

Let's consider an application of the binomial distribution to genetics.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Example. In mice, an allele A for agouti – or gray-brown, grizzled fur – is dominant over the allele a, which determines a non-agouti color. If an $Aa \times Aa$ cross produces 4 offspring, what is the probability that exactly 3 of these have agouti fur?

From the Mendelian model, we know that, for any particular offspring, the probability of the agouti phenotype is $3/4$. Although it is tempting to leap to the conclusion that this means 3 of the 4 offspring must have agouti fur, that is in fact incorrect.

We will first compute the probability that 3 of the 4 progeny have agouti fur without using the binomial distribution, working from the basic laws of probability instead. If we let A represent an agouti offspring and N non-agouti, then there are four ways in which exactly 3 of the 4 offspring could have agouti fur: in order of birth, the offspring might be $NAAA$, $ANAA$, $AANA$, or $AAAN$. Thus, we can use the multiplication and addition rules of probability to find

$$\begin{aligned}
&\mathcal{P}(\text{exactly 3 of 4 offspring has agouti fur}) \\
&\quad = \mathcal{P}(NAAA) + \mathcal{P}(ANAA) + \mathcal{P}(AANA) + \mathcal{P}(AAAN) \\
&\quad = \frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} \\
&\quad = 4\cdot\frac{27}{256} = .421875.
\end{aligned}$$

Now let's redo the computation using the binomial distribution. We'll call having agouti fur a success, so $p = 3/4$, $q = 1/4$. Then,

$$\begin{aligned}
\mathcal{P}(\text{exactly 3 of 4 offspring has agouti fur}) &= \binom{4}{3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right) \\
&= \frac{4!}{3!1!}\left(\frac{27}{256}\right) = .421875.
\end{aligned}$$

► If you decide to call having non-agouti fur a success, that changes the details of the work in this computation, but not the answer. How do the details change?

Notice that even though each offspring of the $Aa \times Aa$ cross has a $3/4 = .75$ chance of having the agouti phenotype, the probability that exactly 3 of 4 offspring have the phenotype is considerably lower, at around .42.

► Why is this statement not contradictory?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Since Mendel's studies focused on *diallelic* genes, that is, genes with exactly two alleles, the binomial distribution very naturally fits this setting. Of course, many genes have more than two alleles. Nonetheless, by grouping alleles into two categories – healthy and diseased, or dominant and recessive – the binomial distribution can often be used to make genetic predictions even when more alleles exist. For instance, in the agouti fur example, the symbol a actually represents a number of different alleles, each associated with different fur colorings and patterns, but all recessive to agouti. Because we are only concerned with the agouti phenotype, we can lump all others together in our analysis.

Example. What is the probability that exactly 4 of 10 mice from an $Aa \times Aa$ cross have agouti fur?

We'll use the same setup as before, only this time we are interested in determining the probability of $k = 4$ successes (agoutis) in $n = 10$ trials (births). This probability is

$$\begin{aligned}\mathcal{P}(\text{4 agouti in 10 births}) &= \binom{10}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^6 \\ &= \left(\frac{10!}{4!6!}\right)\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^6 \approx .01622.\end{aligned}$$

We can use the binomial distribution together with the addition rule to solve even more difficult problems.

Example. What is the probability that more than half of six progeny of a $Aa \times Aa$ cross have the agouti phenotype?

Continuing to think of an offspring with agouti fur as a success, we need to calculate $\mathcal{P}$(at least 4 successes in 6 trials). But this is the same as

$$\begin{aligned}&\mathcal{P}(\text{4 successesin 6 trials}) + \mathcal{P}(\text{5 successes in 6 trials}) \\ &\quad + \mathcal{P}(\text{6 successes in 6 trials}) = \binom{6}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^2 \\ &\quad + \binom{6}{5}\left(\frac{3}{4}\right)^5\left(\frac{1}{4}\right)^1 + \binom{6}{6}\left(\frac{3}{4}\right)^6\left(\frac{1}{4}\right)^0 \approx .83057.\end{aligned}$$

Thus, it is quite likely that more than half of the six offspring have agouti fur.

Let's return to considering the number of agouti mice in a cross producing four progeny. Similar computations to those above give the probabilities that

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.5. *Probabilities of Exactly i Agouti Mice Among Four Progeny of Aa × Aa Cross*

i	0	1	2	3	4
$\mathcal{P}(i)$	.00390625	.046875	.2109375	.421875	.31640625

exactly 0, 1, 2, 3, or 4 of the 4 progeny have agouti fur, as shown in Table 6.5. Of course, the entries in this table add to 1.

While Table 6.5 tells us the probability of any outcome of this four-progeny mouse cross, a useful summary of the table is the *expected value* of the number of agouti mice progeny. Informally, the expected value tells us how many agouti progeny we might expect when four offspring are produced. You should think of the expected value as an average of the outcomes, with each outcome weighted by the probability it occurs. To be more precise,

Definition. For any probability distribution describing a random variable with a finite number of possible outcome values i, the *expected value* is defined as

$$E = \sum_{\text{outcomes } i} i \cdot \mathcal{P}(i).$$

Because it is an average weighted by probabilities, the expected value of a random variable might not be an integer, even if the random variable can have only integer outcomes.

For the example described by Table 6.5, we find

$$E = 0 \cdot .00390625 + 1 \cdot .046875 + 2 \cdot .2109375 + 3 \cdot .421875 + 4 \cdot .31640625 = 3.$$

Thus, in this example, the expected value seems to be capturing our naive belief that 3 of the 4 mice should have agouti fur, since the probability that any particular mouse does is 3/4.

As you will see in the exercises, whenever a random variable with a binomial distribution is used, something similar happens. More specifically, the expected value for the number of successes in n trials, assuming the probability of any one success is p, is

$$E = np.$$

This should seem reasonable, because it simply states that if for each trial the fraction of times you get a success is p, then of n trials, you expect np successes.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

► What is the expected number of agouti offspring in 10 births?

Expected values for two random variables have a nice additive property. Suppose for the mouse cross above, we consider the random variables

$$X_1 = \text{the no. of agouti mice in a litter of 4,}$$
$$X_2 = \text{the no. of agouti mice in a litter of 5.}$$

Then $X_1 + X_2 =$ the no. of agouti mice in a litter of 9, since we can think of the first 4 births and the last 5 births as two separate groups. In this case, it is easy to check that

$$E(X_1 + X_2) = E(X_1) + E(X_2), \tag{6.2}$$

because the left-hand side is $9(3/4)$, and the right-hand side is $4(3/4) + 5(3/4)$. In fact, Eq. (6.2) holds for any two random variables, as you will see in the exercises.

The χ^2 distribution. Although the binomial distribution is useful in computing the probabilities of certain types of outcomes in repeated trials, many other probability distributions arise in biology. A particular useful distribution for genetics is the χ^2 distribution. Rather then predicting the likelihood of certain outcomes, the main use of the χ^2 distribution is to determine whether the outcome of an experiment fits a particular probabilistic model.

For instance, the Mendelian model predicts that all the phenotypic ratios in Table 6.2 should be 3:1. However, none of them were exactly 3:1, even though they were close. With Mendel's data the results are so close to 3:1 that few would doubt the model applies, but what if they had been further from that ratio? How far could they deviate before we might doubt the model?

In designing an experiment, a scientist ideally has a *hypothesis* to test. For Mendel's experiment, this might be: *The principal of segregation, that parents pass on each of their alleles to progeny separately and with equal likelihood, holds.* This hypothesis implies that a cross between F_1 hybrids should yield a phenotypic ratio of approximately 3:1. The larger the number of offspring, the closer we expect the experimental ratio to match the theoretical 3:1.

If data collected from the experiment is in line with the expected results, then evidence has been gathered in support of the hypothesis. If the data deviates a great deal from the expected values, then a scientist must reconsider the validity of the hypothesis; perhaps the hypothesis was wrong, or perhaps the experiment was poorly designed. An important issue for the researcher, then, is how to decide whether the data *fits* the hypothesis.

Table 6.6. *Progeny of* $Gg \times Gg$

Phenotype	Observed No.	Expected No.
Yellow seeds	231	245.25
Green seeds	96	81.75
TOTAL	327	327

The χ^2-statistic is one way to measure *goodness of fit* of data to the hypothesis in an experiment. From the data and hypothesis, we compute a certain number according to a formula given below and denote it by χ^2. If this χ^2-statistic is large, the fit is poor. If it is small, the fit is good. An understanding of the probability distribution for this particular random variable – the χ^2-distribution – will allow us to decide how large χ^2 must be for us to consider it unlikely that the hypothesis is correct.

To illustrate how the χ^2-statistic is used, let's apply it to one of Mendel's experiments, to test the hypothesis that the principal of segregation applies to seed color. (This was one of Mendel's hypotheses, though he did not phrase it this way.) In the laboratory, we cross hybrids $Gg \times Gg$ and obtain 327 progeny. Under our hypothesis, we expect that 3/4 of these, $(.75)(327) = 245.25$, will be phenotypically dominant with yellow seeds, and the remainder, $(.25)(327) = 81.75$, will have green seeds. The experiment turns out to produce data that is a bit off from that, as shown in Table 6.6.

The χ^2-statistic is defined as

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

Here, O_i and E_i denote the observed and expected frequencies, that is, the observed and expected numbers as in Table 6.6. Each expression $(O_i - E_i)$ measures deviation of an observation from what we expect, and because this expression is squared, any chance of positive terms canceling with negative ones is eliminated. Dividing each term by E_i gives us a sense of how large the deviation is relative to the expected number. Summing gives us a measure of total deviation.

In this experiment, we have $n = 2$ classes and find

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(231 - 245.25)^2}{245.25} + \frac{(96 - 81.75)^2}{81.75} \approx 3.312.$$

If χ^2 were smaller, we would know the data fit the hypothesis better, and if

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

it were larger, the fit would be worse. We still don't know whether 3.312 is small enough to consider the fit to be good.

Before proceeding, there is one other issue we must understand about χ^2-statistics. Because χ^2 involves adding up a number of positive terms, we would expect its value to be larger whenever there are more terms. This is captured in the idea of a parameter called the *degrees of freedom*. Counting the degrees of freedom can be quite difficult, but a rule of thumb is that there is one degree of freedom for each class whose size can vary freely. In this example, if we imagine the size of the first class (the yellow seed phenotype) varies freely (it could be any number from 0 to 327), then the size of the second class (the green seed phenotype) is obtained by subtracting the first from the total 327. This means we have one degree of freedom. More generally, if we had n classes in a test, then the first $n - 1$ of them could range freely, but the last is constrained. This corresponds to $n - 1$ degrees of freedom. The more degrees of freedom in a test, the larger you might find the χ^2-statistic to be, because it requires summing more positive numbers. To judge the size of a particular χ^2-statistic, we must take this into account.

With the degrees of freedom specified, statisticians have studied the χ^2 distribution. Although a formula for the distribution is too complicated to give here, information from it is incorporated in tables and in software. This makes it possible to compute, for a specified number of degrees of freedom, the probability that the χ^2-value lies in any specified range, assuming the hypothesis holds.

Keep in mind that, even when the hypothesis is true, every time we do an experiment, we will get different data and a different χ^2-statistic describing the fit. Most of these will be small, but some will be large because of chance. We would like our goodness-of-fit test to be flexible enough to accommodate this variation. So, to decide whether we consider our value of χ^2 to be too large for the data to fit the hypothesis, we pick a *significance level*, for instance $\alpha = .05$. This means we decide to view χ^2 as too large if the probability of getting a lower value is at least $1 - \alpha = 95\%$ when the hypothesis is true.

If we consult a table, such as the abbreviated Table 6.7 at the end of this section, we find that the critical value for a χ^2-statistic with one degree of freedom at the .05 level of significance is $\chi^2_{\text{critical}} = 3.841$. This means that, assuming the hypothesis is correct, only 5% of the time would we calculate a value of χ^2 that was 3.841 or larger. Thus, if our statistic is larger than 3.841, we say the data do not support our hypothesis at the .05 level of significance. However, if our statistic is less than 3.841, we find that the data do support the hypothesis at the .05 level of significance.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Since for Mendel's experiment, $\chi^2 = 3.312 < 3.841 = \chi^2_{\text{critical}}$, the value of χ^2 is not too large, and the experiment supports the hypothesis that the alleles for seed color segregate.

If, instead, our statistic had turned out to be larger than χ^2_{critical}, leading us to reject the hypothesis at the .05 level, a number of things could be responsible. It could be that the hypothesis was wrong (exactly what χ^2-statistics are trying to test), or it could be that our hypothesis is correct and we just happened to obtain extreme data through randomness.

In fact, even when the hypothesis is actually correct, this second case will happen 5% of the time. If we breed pea plants that are perfectly described by the Mendelian model again and again, as in this experiment, and calculate χ^2-statistics for each of these trials, then about 5% of the time we would expect to see χ^2-values larger than χ^2_{critical}. A χ^2 test is not capable of definitively telling us whether the hypothesis is true or not.

- Sometimes critical values corresponding to a level of .01 or .1 are used. Which of these makes it more likely that you will doubt the hypothesis you are testing? Which level insists on a closer fit of the data to the expected frequencies?

A significance level of .01 means that we only consider a χ^2-value to show a poor fit if it is larger than what would occur 99% of the time when the hypothesis is true. That means we are less likely to reject the hypothesis erroneously. On the other hand, the significance level .1 insists on a closer fit for us to feel the data supports the hypothesis. With $\alpha = .1$, we are more likely to reject the hypothesis erroneously.

As you can probably imagine, we are just at the tip of the iceberg in discussing χ^2-statistics. There is much more to learn about them as they are used ubiquitously in the scientific world. You will get some practice in the exercises, but a course in statistics is really necessary to delve deeper.

Table 6.7. χ^2_{critical} *Values at Significance Level* α

d.f.	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
1	2.70554	3.84146	6.63490
2	4.60517	5.99147	9.21034
3	6.25139	7.81473	11.3449
4	7.77944	9.48773	13.2767
5	9.23635	11.0705	15.2767

Note: d.f. denotes degrees of freedom.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Problems

6.2.1. List all the ways that you might have exactly 3 heads among 5 coin flips. Then compute $\binom{5}{3}$ by Eq. (6.1) to verify that it gives the correct count.

6.2.2. In the text, the binomial distribution is used to find the probability of exactly 1 of 3 coin flips producing tails. Find the probabilities of exactly none, exactly two, and exactly three tails in this situation. What is the sum of these four probabilities?

6.2.3. Verify the entries in Table 6.5.

6.2.4. Use a calculator or computer to find $\binom{10}{k}$ for each $k = 0, 1, 2, \ldots, 10$. A MATLAB command to do this type of calculation is `nchoosek(10,0)`.

a. For which k is $\binom{10}{k}$ smallest? For these particular values of k, explain why it has the value it does by thinking in terms of choosing objects.

b. For which k is $\binom{10}{k}$ largest? Is this intuitively reasonable? Explain.

c. What patterns do you notice in your calculations? Do the patterns hold if 10 is replaced by other numbers?

6.2.5. Explain the following results not by referring to formula (6.1), but in terms of choosing objects.

a. $\binom{n}{1} = n$ and $\binom{n}{n-1} = n$ for any n.

b. $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any n.

6.2.6. Suppose a family has six children.

a. What is the probability that four are boys and two are girls?

b. Give the probability distribution (i.e., the seven probabilities) that the family has $0, 1, \ldots,$ or 6 boys. How would your answer change if you were to list the probability distribution for the number of girls in the family?

c. What is the expected number of boys in the family?

d. What is the probability that the family has four or more girls?

6.2.7. In the text, the binomial distribution is used to find the probability that exactly 3 of 4 offspring have agouti fur from a cross of mice heterozygous for agouti fur.

a. Find the probabilities that exactly 30 of 40 offspring of this cross have agouti fur. Then, find the probability that exactly 300 of 400 offspring have agouti fur.

b. Can these results be consistent with the fact that, in a large number of such offspring, we would expect 3/4 of them to have agouti fur? Explain.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

6.2.8. If you roll a fair die once, what is the expected value of the outcome?

6.2.9. Suppose you roll two fair dice and add the results.

a. Calculate the expected value of the outcome by first finding the probabilities of each of the outcomes $2, 3, 4, \ldots, 12$, and then computing a weighted average of the outcomes.

b. Let X_1 and X_2 be random variables denoting the outcome of the roll of the first and second die, respectively. Find $E(X_1)$, $E(X_2)$, and $E(X_1 + X_2)$.

6.2.10. When using the binomial distribution in applications, it does not matter which of the two trial outcomes you consider a success. Use the binomial distribution to calculate the probability of 10 rolls of a die producing three sixes as follows.

a. If you call "producing a six" a success, what should p, q, n, and k be in the binomial formula for this probability? What is the resulting probability?

b. If you call "not producing a six" a success, what should p, q, n, and k be in the binomial formula for this probability? What is the resulting probability?

6.2.11. Part of the reason the formula for the binomial distribution gave the same result in both parts of the last problem was because $\binom{n}{k} = \binom{n}{n-k}$.

a. Explain in intuitive terms, in terms of choosing k or $n - k$ objects from n objects, why this formula should hold.

b. Explain why the mathematical formula (6.1) shows this formula holds.

6.2.12. One form of albinism (lack of pigment) in humans is caused by a recessive allele a. Suppose an homozygous albino marries a heterozygote, and the couple has two children.

a. What is the probability their first child will be an albino?

b. What is the probability their first child will be an albino and their second child will have normal skin pigment?

c. What is the probability exactly one of their two children will be an albino?

d. What is the probability at least one of their two children will be an albino?

e. What is the expected number of their children that will be albino?

6.2.13. Mice homozygous for a recessive allele, f, are fat. Suppose a *dihybrid* cross, $AaFf \times AaFf$, is carried out by experimenters. Here, A denotes the agouti allele.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

a. How many of 25 progeny are expected to be fat with agouti fur?
b. What is the probability that exactly 4 of 25 progeny will be fat with agouti fur?
c. What is the probability that, at most, 4 of the 25 progeny will be fat with agouti fur?
d. What is the probability that at least 4 of the 25 progeny will be fat with agouti fur?

6.2.14. In a certain population of rats, the probability of an individual surviving through its first year is .5. For the rats who make it to age one, the probability of surviving a second year is .25, and for those who make it to age two, the probability of surviving a third year is also .25. All rats die before the end of their fourth year.
a. What is the probability that the age (in years, rounded down) at death of a rat is 0? 1? 2? 3? Why should these add to 1?
b. What is the expected age at death of one of these rats?

6.2.15. The yellow-lethal allele is dominant for yellow fur color, but lethal to homozygous embryos. Suppose two mice, both heterozygous for the yellow-lethal mutation, are crossed and produce 12 viable progeny.
a. What is the probability that exactly five of them will have normal coloring?
b. What is probability that 10 or more of the progeny will be yellow?
c. What is the probability that at most three of the progeny will be yellow?

6.2.16. In humans, the hereditary Huntington disease is caused by a dominant mutation. Onset of Huntington disease occurs in midlife, between 35 and 44 years of age typically, and the progressive disorder leads eventually to death. Suppose, in a married couple, one individual carries the allele for Huntington disease. They have four children.
a. What is the probability that none of their children will develop Huntington disease?
b. What is the probability that at least one of their children will develop Huntington disease?
c. What is the probability three or more of their children will develop Huntingdon disease?

6.2.17. In a trihybrid cross, $AaBbCc \times AaBbCc$, what is the probability that exactly 20 of 30 progeny will display the dominant phenotype for all three traits? What is the probability that at least two of the progeny will display the dominant phenotype for at least one of the traits?

6.2.18. The goal of this problem is to derive formula (6.1) for counting combinations. Formally, a combination of n things taken k at a time is

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

an *unordered* k element subset of a set of n elements. However, it's better to think of it more concretely, as follows. Imagine a box of n balls with the numbers $1, 2, 3, \ldots, n$ printed on them. You pick out k of these balls and first place them in a row, in the order you picked them. Then, since you don't care about the order, you dump them in a bag. That's a combination. The number of different bags of balls you might end up with is $\binom{n}{k}$.

a. When you pick the first ball out of the box, how many different choices could you make for it? When you pick the second ball, why are there only $n-1$ choices for it? For the lth ball, why are there $n-l+1$ choices?

b. Why does part (a) indicate that, when the k balls are all in a row, there are $n(n-1)(n-2)\cdots(n-k+1)$ possible choices you might have made? (The count of these ordered choices is sometimes called a *permutation*.)

c. Several different ordered choices might lead to the same collection of balls in the bag, so the answer in part (b) is bigger than the number of combinations. To see how much bigger, it's easiest to imagine having the balls in the bag, and (going backward in time) putting them back in some order in a row. Using reasoning similar to parts (a) and (b), explain why there are $k(k-1)\cdots 2\cdot 1 = k!$ choices of ways this could be done.

d. Using parts (b) and (c), conclude $\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$.

e. Explain why this formula can also be written as formula (6.1).

6.2.19. The binomial distribution received its name because of a relationship to the expression $(x+y)^n$, a power of a binomial. In fact, the numbers $\binom{n}{k}$ are often called the *binomial coefficients*, because they give the coefficients in the expansion of $(x+y)^n$. That is,

$$(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n}y^n. \qquad (6.3)$$

a. Check this for $n = 2$, 3, and 4, using Eq. (6.1).

b. By thinking of $(x+y)^n$ as a product of n copies of $(x+y)$, explain why this product will produce a term $x^k y^{n-k}$ for each way we can choose k of the copies. Explain why this justifies formula (6.3).

c. What is the sum $\sum_{i=0}^{3}\binom{3}{i}$?, $\sum_{i=0}^{4}\binom{4}{i}$? Give a formula for $\sum_{i=0}^{n}\binom{n}{i}$.

6.2.20. Suppose a trial has probability of success p, so the number of successes in n trials is described by the binomial distribution. Show the expected value for the number of successes in n trials is $E = np$ as

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

follows:

a. Express the expected value as a sum involving factorials and powers of p and q.

b. Show

$$i\frac{n!}{(n-i)!i!}p^i q^{n-i} = pn\frac{(n-1)!}{(n-i)!(i-1)!}p^{i-1}q^{(n-1)-(i-1)}.$$

c. Use part (b) to factor pn from your expression in part (a). Then use Eq. (6.3) to complete the problem.

6.2.21. The goal of this problem is to show that expected values of random variables are additive, as claimed in Eq. (6.2). Only a special case will be considered, where X_1 and X_2 are two independent random variables that, for simplicity, can take on only integer values between 1 and N.

a. Explain why the expected value of $X_1 + X_2$ is

$$E(X_1 + X_2) = \sum_{i=1}^{N}\sum_{j=1}^{N}(i+j)\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j).$$

b. Through algebra, show this can be written as

$$\sum_{i=1}^{N} i\mathcal{P}(X_1 = i)\sum_{j=1}^{N}\mathcal{P}(X_2 = j) + \sum_{j=1}^{N} j\mathcal{P}(X_2 = j)\sum_{i=1}^{N}\mathcal{P}(X_1 = i).$$

c. What are $\sum_{i=1}^{N}\mathcal{P}(X_1 = i)$ and $\sum_{i=1}^{N}\mathcal{P}(X_1 = i)$? Use this to conclude Eq. (6.2) holds.

6.2.22. Suppose an $Aa \times Aa$ cross produces 1,000 progeny, N with the dominant phenotype, and $1{,}000 - N$ with the recessive phenotype.

a. For $N = 700$, compute the χ^2-statistic to test whether this data fits the Mendelian model. Using a significance level of $\alpha = .05$, is the data in accord with the model?

b. Repeat part (a) with $N = 725$.

c. What is the smallest value of N that would be judged in accord with the model (at the $\alpha = .05$ level)? The largest value of N?

6.2.23. Explain informally why in Table 6.7, the entries get larger as you move across the rows. Explain informally why they get larger as you move down the columns.

6.2.24. The data in Table 6.8 is from Mendel's experiments with genes for seed shape and color resulting from $WwGg \times WwGg$ crosses ($W =$

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.8. *Progeny of* $WwGg \times WwGg$

Phenotype	Observed No.
Round, yellow	315
Round, green	108
Wrinkled, yellow	101
Wrinkled, green	32

round; w = wrinkled; G = yellow; g = green). Use χ^2 to test if the genes for seed color and shape assort independently in pea plants. Because there are four phenotypes, there are $4 - 1 = 3$ degrees of freedom.

6.2.25. The critical value of a χ^2-statistic comes from a theoretical χ^2 distribution with appropriate number of degrees of freedom. Figure 6.2 shows a graph of a typical χ^2 distribution.

In such a graph, the values of χ^2 are along the horizontal axis, and probabilities of χ^2 falling in any interval are represented by the area above that interval and below the curve. The total area between the curve and the horizontal axis is 1 unit and corresponds to 100% or a probability of 1. The critical value χ^2_{critical} at significance level α is the value on the horizontal axis that leaves an area of α to the right. In Figure 6.2, this area is shaded for $\alpha = .05$.

a. Suppose you are performing a χ^2-test and choose a significance level of .01 (or 1%). Where, approximately, would the critical value fall on the horizontal axis in Figure 6.2?

b. Notice that the bulk of the area under the curve is just a bit to the right of the vertical axis and there is very little area under the right

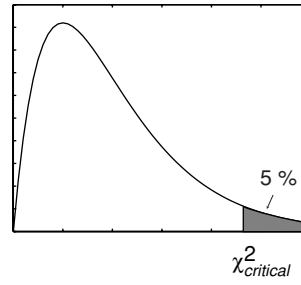

Figure 6.2. χ^2-Distribution.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

tail of the curve. If your data is well explained by your experimental hypothesis, where do you expect your calculated χ^2-statistic to fall? How is the shape of this curve related to the goodness-of-fit test?

6.3. Linkage

After receiving little attention for more than 30 years, Mendel's theory of inheritance eventually became well-accepted through the efforts of the British geneticist William Bateson and others. Since Mendel only hypothesized the existence of genes, it was necessary to find the physical basis of these units of inheritance. Around the turn of the century, biologists suspected strongly that genes resided on chromosomes, large thread-like structures that could be stained and viewed under a microscope during cell division. Evidence for this was given by the American geneticist Thomas Hunt Morgan in 1910 and his coworkers. Morgan's group worked with fruit flies, *Drosophila melanogaster*, a favorite of geneticists, since they reproduce quickly and in great abundance and have some readily observable traits with simple variants.

Sex-linked genes. Let's consider one of Morgan's experiments to see how his laboratory was able to discover the important role played by chromosomes in inheritance. After 2 years of breeding *Drosophila*, a mutant male fruit fly with white eyes was born. (Normal, or *wildtype*, eye color is red.) This white-eyed male was crossed with wildtype red-eyed females, and the resulting F_1 generation had all red eyes, indicating that the new mutant allele was recessive. Then, the F_1 generation was interbred to produce F_2.

► Assuming Mendel's model applies, what fraction of the F_2 population should have white eyes?

The basic model predicts that, regardless of sex, $1/4$ of F_2 would be homozygous recessive, and hence have white eyes. However, when the F_1 generation were intercrossed, the F_2 were observed with phenotypes as given in the middle column of Table 6.9. In a striking departure from the expected values, there is a total absence of any white-eyed females. Also, roughly half of the males had white eyes, rather than the predicted $1/4$. While roughly $1/4$ of all progeny had white eyes, they are not distributed equally among the sexes. This is strong evidence for a connection between the determination of sex and the behavior of the eye color gene.

In a second experiment in Morgan's laboratory in which a female from F_1 was crossed with the mutant male, phenotypes of (very) roughly equal

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.9. *Progeny of Two Crosses*

Phenotype	$F_1 \times F_1$	$F_1 \times$ mutant
Red-eyed, female	2459	129
White-eyed, female	0	88
Red-eyed, male	1011	132
White-eyed, male	782	86
TOTAL	4252	435

frequency occurred, as shown in the last column of Table 6.9. However, this data is *not* in contradiction with Mendelian genetics, since a $Ww \times ww$ cross would produce equal numbers of each phenotype, regardless of sex.

The first experiment points out the need for a new model consistent with its outcome. However, any new model must be capable of predicting the outcome of the second as well.

At about the same time that Morgan concluded from these experiments that the inheritance of eye color must somehow be related to the determination of sex, he also noticed a relationship between sex and chromosomes under microscopic inspection of the flies' cells. Although all chromosomes came in matching pairs in female *Drosophila*, male *Drosophila* had one nonidentical pair of chromosomes. Moreover, one of the chromosomes from the nonidentical pair in males was morphologically identical to a pair in females. Morgan suspected that this set of chromosomes, the *sex chromosomes*, must control sex determination in fruit flies and that a gene for eye color must lie on this chromosome pair.

Morgan proposed a model for this sex-linked gene behavior that used chromosomes to explain the observations from experimental data. We denote the identical sex chromosomes in females by XX, and the corresponding differing chromosomes in males by XY. In addition, we'll use w to denote the white-eye allele, and w^+ the wildtype red-eye allele. (Such notation is common for the wildtype alleles of any gene.) Hypothesizing that the eye-color gene lies on the X chromosome only, we let X^w denote a sex chromosome carrying the white-eye allele, and X^+ one carrying the wildtype allele.

In Morgan's initial experiment, the females were genotype X^+X^+ and the mutant male X^wY. Now, assuming *segregation of chromosomes* in gamete formation, in the F_1 generation we expect equal numbers of the genotypes X^+X^w and X^+Y. We continue to view the white-eyed mutation as recessive, so each female will have red eyes due to the X^+X^w genotype. Similarly, each of the F_1 males carries only a wildtype allele X^+ so they also have red eyes. The presence of the gene for eye-color on the X chromosome, with no

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.10. *Punnett Square for* $X^+X^w \times X^wY$

	X^+	X^w
X^w	X^+X^w	X^wX^w
Y	X^+Y	X^wY

corresponding gene on the Y chromosome is consistent with Morgan's data on the phenotypic make-up of F_1.

We will leave analysis of the experiment leading to the data in the middle column of Table 6.9 to the exercises and instead consider Morgan's second experiment. When F_1 females were crossed with the mutant male, Morgan was crossing heterozygous females X^+X^w with *hemizygous* males X^wY. Again, assuming segregation of chromosomes, the results of this cross are shown in the Punnett square of Table 6.10.

Now, each genotype in the table is equally likely for progeny, and each genotype gives a different phenotype. In the top row, the phenotypes are red-eyed female and white-eyed female; in the bottom row, they are red-eyed male and white-eyed male. This corresponds roughly with the approximately equal numbers in the last column of Table 6.9. (In the exercises, you are asked to perform a χ^2-test to test more rigorously if the hypothesis of X-linked inheritance of eye color meshes well with this data.)

Because males and females have a different number of X chromosomes, X-linked traits are often manifested in different proportions in the two sexes. For a female *Drosophila* to have white eyes, she must be homozygous for the mutant allele, X^wX^w, receiving a (possibly rare) mutant allele from each parent. However, for a male to have white eyes, he needs only one mutant allele so that his genotype is X^wY. As a consequence, recessive X-linked traits are more likely to appear in males. In humans, certain types of color blindness, hemophilia, and mental retardation from fragile X syndrome are X-linked traits that are found almost exclusively in males.

Linked genes and genetic mapping. While sex-linked genes required a modification of the Mendelian model, other experiments from Morgan's laboratory pointed to additional problems with the idea of independent assortment of genes. Even when sex determination was not involved, numerous examples were found of data inconsistent with that assumption.

One such example concerns two genes in *Drosophila*. One gene affects wing shape, with the dominant allele causing straight wings and the recessive

Table 6.11. *Progeny of Cross*

Phenotype	No.
Straight wings, red eyes	520
Straight wings, purple eyes	133
Curved wings, red eyes	129
Curved wings, purple eyes	467
TOTAL	1,249

causing curved wings. The second gene affects eye color, with the dominant allele causing red eyes and the recessive causing purple eyes. Crossing a homozygote recessive for both genes (with curved-wing, purple-eye phenotype), with a heterozygote for both genes (with straight-wing, red-eye phenotype), produces data like that in Table 6.11.

► If the two genes assort independently, what is the expected phenotypic ratio? Is the data in line with that?

Although the basic Mendelian model, with independent assortment of the two genes, would have predicted that all four phenotypes were equally likely, the data show clear deviation from this. The inheritance of the two genes seems to be linked, in that there is a definite tendency for the progeny to have a phenotype similar to one or the other of the parents.

This linkage comes from the relationship of genes to chromosomes and the manner in which gametes are formed. The *chromosomal theory of heredity* revised and improved the Mendelian model by taking into account the physical location of genes on chromosomes and modeling such linkage.

Most cells in diploid organisms contain a set of pairs of chromosomes, with one chromosome in a pair inherited from each parent. Chromosomes are divided into two types: *autosomes* (nonsex chromosomes) and *sex chromosomes*. Chromosome number varies greatly between species and seems in no way to reflect developmental complexity; humans have 46 chromosomes, *Drosophila* 8, and cats 72.

According to the chromosomal theory of heredity, gametes are formed by the segregation of chromosomes into reproductive cells, rather than the simpler segregation of genes that Mendel imagined. Genes reside on chromosomes, arranged in a linear fashion. Somatic cells, or body cells, are diploid in that they contain the full count of $2n$ chromosomes in a species. Gamete cells have only half the number of chromosomes, n, and are called *haploid*. At fertilization, two gametes (e.g., an egg and sperm) are united to form a zygote, from which a new diploid offspring develops.

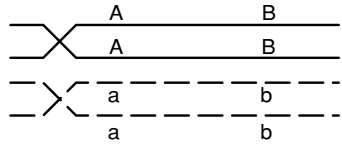

Figure 6.3. Tetrad before crossing over; four chromatids are visible.

When gametes are formed, they do not simply receive a copy of one of the chromosomes in each pair. Instead, a complicated and not completely understood process of *crossing over* provides a source of *genetic recombination*. The chromosome passed along to the gamete is not an identical copy of either one of the parental chromosomes, but instead an amalgam of the parental chromosome pair, with some genes from each.

► If no crossing over occurred, how would the principle of independent assortment of genes need to be modified? If two genes were on different chromosomes, would they assort independently? What if they were on the same chromosome?

Let's look more closely at how crossing over works. In the process of gamete formation, chromosomes replicate forming identical *chromatids* joined at a *centromere*. Next, matching chromosomes gather together and form *homologous pairs*. This arrangement, known as a *tetrad*, can be seen in Figure 6.3.

In crossing over, two chromatids in the tetrad exchange genetic material. If the chromatids belong to different chromosomes, then this might result in an exchange of alleles. For example, suppose the solid chromosome in Figure 6.3 was inherited from the mother and contains dominant alleles for two genes AB, and that the dashed chromosome, inherited from the father, has recessive alleles for these genes, ab. (Note that the individual in this example is heterozygous for these two genes, $AaBb$.)

During crossing over, two chromatids swap DNA as shown in Figure 6.4. Since nonidentical chromatids are involved in crossing over, they exchange alleles B and b for the second gene. The *parental types* AB and ab occur in the tetrad before crossing over, but after crossing over four genotypes are represented: AB, Ab, aB, and ab. The two new genotypes, Ab and aB, the results of crossing over, are *recombinants*. In the final steps of gamete formation, the four chromatids separate, with each one going into a different gamete.

Because it is so important biologically, we point out again that only two of these gametes are identical to a parental chromosome; the two recombinant

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

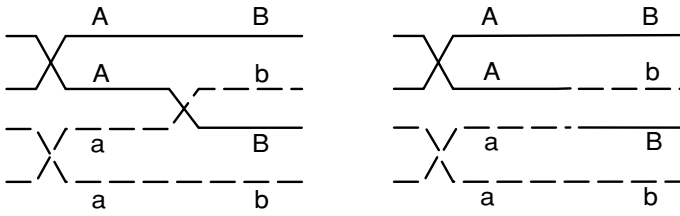

Figure 6.4. Tetrad during (L) and after (R) crossing over.

gametes, if they ultimately unite with other gametes and develop into full organisms, will introduce new genetic combinations into the population. In fact, more than one crossover can occur between homologous chromosomes, so tremendous possibilities for genetic variation are introduced. New variation is of course the raw material for evolution, since recombinants may be better adapted for survival and reproduction.

From a modeling point of view, the behavior of genes on a chromosome during gamete formation can now be captured by the probability of a crossover occurring between them. If this probability is low, then alleles for the two genes will tend to be inherited together, and parental types will dominate in the progeny. If this probability is high, then recombinants will be more common in the progeny. A probability of .5 for a crossover, so that the genes essentially behave as if on different chromosomes, would result in independent assortment. Any divergence from independent assortment is known as *linkage*.

Alfred Sturtevant, who at the time was an undergraduate student working in Morgan's laboratory, realized that the observed frequencies of crossovers could be used to create a genetic map. If we imagine that a chromosome is a long string with genes ordered along it, then it seems natural to expect that, for any little piece of the chromosome, there is some specific probability of a crossover occurring there. Sturtevant's idea was that this probabilistic behavior could be used to give an abstract notion of genetic distance, and then from that distance a map could be constructed. Specifically, he defined the *genetic distance* between two genes on a chromosome as the average number of crossovers that were observed between them during formation of many gametes. If between two genes crossovers are rare, the distance between them is small; if many crossovers typically occur, the distance is large.

Notice that genetic distance is statistical in nature. More precisely, for any stretch of a chromosome, there is a random variable giving the number of crossovers that occur on that piece in gamete formation. Its probability distribution describes the chance of 0, 1, 2, . . . , crossovers occurring in that piece.

The expected value of this random variable (or more simply, the expected number of crossovers) is what Sturtevant's average was estimating. Because expected values are additive by Eq. (6.2), they will behave just like distances on a map. We formalize these ideas with a definition.

Definition. The *genetic distance* or *linkage distance* between two genes on a chromosome is the expected number of crossovers that occur between the genes in gamete formation.

Because the expected value is an average number of crossovers, theoretically a genetic distance could take on any value from 0 upward. For physically close genes, genetic distances will tend to be small, since crossovers are less likely to occur, whereas for physically distant genes, distances will tend to be larger. The type of map we will construct from crossover data is called a *linkage* or *genetic map*. This map will show the linear arrangement of the genes on a chromosome, with genetic rather than physical distances separating genes.

Let's see how a *two-point testcross* can place two genes on a linkage map. Suppose we suspect that, in *Drosophila*, the genes for curved wings *c* and purple eyes *pr* are linked. For genotypes of linked genes, we use a special notation to keep track of which alleles are on which chromosome in a given pair. For instance, we write $c\ pr/c\ pr$ for a homozygous recessive *Drosophila*, where the slash separates alleles inherited from different parents. There are now several different ways a fly could be heterozygous at both genes; $c\ pr/c^+\ pr^+$ and $c^+\ pr/c\ pr^+$ are different configurations.

As a first step in genetic mapping, we cross true-breeding, curved-wing, purple-eyed *Drosophila* with true-breeding wildtype flies: $c\ pr/c\ pr \times c^+\ pr^+/c^+\ pr^+$. Notice that all the progeny in F_1 are genotypically $c^+\ pr^+/c\ pr$ and phenotypically wildtype, since curved wings and purple eyes are recessive traits.

Next, we cross F_1 flies with curved-winged, purple-eyed flies to produce F_2. This testcross is $c^+\ pr^+/c\ pr \times c\ pr/c\ pr$, and we suppose that the data in Table 6.11 came from such an experiment. As we noticed before, there is a discrepancy between the data and the numbers predicted by Mendelian genetics. Moreover, because there are two large phenotypic classes that resemble the parents – red-eyed with straight wings and purple-eyed with curved wings, and two smaller nonparental phenotypic classes – there is evidence for linkage.

- What are the possible genotypes of the F_2 progeny in this second cross? Which of these are parental types and which are recombinants?

Notice how this testcross was designed to test for linkage and crossing over. In the doubly recessive homozygous parent, crossing over may occur between identical chromatids, but it has no effect on the genotype of the gamete. Such *Drosophila* only create $c\ pr$ gametes. In contrast, the parental-type gametes from the heterozygous parent are $cr^+\,pr^+$ and $c\ pr$, and crossing over results in recombinants $c^+\,pr$ and $c\ pr^+$ that will be phenotypically detectable in progeny.

► Why are the recombinants $c^+\,pr$ and $c\ pr^+$ phenotypically observable in this cross?

Now we can estimate the average number of crossovers that occurred. Because the recombinants $c^+\,pr$ and $c\ pr^+$ result from a crossover, each straight-winged, purple-eyed *Drosophila*, $c^+\,pr/c\ pr$, and each curvy-winged, red-eyed fruit fly, $c\ pr^+/c\ pr$, is the result of a crossover. In the testcross above, we suspect that $133 + 129 = 262$ crossovers took place.

Now, assuming all recombinants were created by a single crossover, the *recombination frequency* (no. of recombinants)/(total no. of progeny) is exactly the same as the average number of crossovers. Thus, the genetic distance is estimated by

$$\frac{\text{no. of recombinants}}{\text{total no. of progeny}} = \frac{262}{1249} \approx .21 \text{ units } \equiv 21\ cM.$$

Genetic distances are usually measured in *centiMorgans* (cM) in honor of Morgan.

In our calculation, we made the assumption that all recombinants were created by a single crossover. What if two crossovers occurred between the genes on a chromatid with $c^+\,pr^+$ and one with $c\ pr$? Then, the gametes produced would be of parental type, and our testcross would produce no evidence of any crossovers (see Figure 6.5). Similarly, if three crossovers occurred between the genes, that would appear to us exactly as if only 1 had occurred. Thus, our use of the recombination frequency may understate the true average. Only if we believe multiple crossovers are very rare between

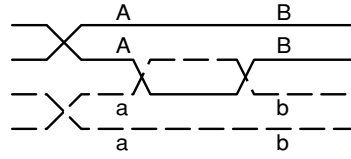

Figure 6.5. A double crossover producing no recombination of genes A and B.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.12. *Progeny of* $gl^+d^+ws^+/gl\ d\ ws \times gl\ d\ ws/gl\ d\ ws$

Phenotype	No.
Normal leaves, tall, normal sheaf	301
Normal leaves, tall, white sheaf	146
Normal leaves, dwarf, normal sheaf	15
Normal leaves, dwarf, white sheaf	1
Glossy leaves, tall, normal sheaf	2
Glossy leaves, tall, white sheaf	17
Glossy leaves, dwarf, normal sheaf	154
Glossy leaves, dwarf, white sheaf	289
TOTAL	925

these genes can we believe our estimate of genetic distance is good. If the genes are close, and the average number of crossovers is small, then our estimation is reasonable.

Now that we have seen how testcrosses can be used as evidence for linkage and for estimating genetic distances, let's extend the method to locating three or more genes on a genetic map. Consider three genes in corn plants with recessive alleles: *d* for dwarf plants, *gl* for glossy leaves, and *ws* for white sheafs. In creating a genetic map, we now have to determine the order of the genes on the chromosome as well as find distances.

To locate the three genes, we make a three-point testcross, $gl^+d^+ws^+/gl\ d\ ws \times gl\ d\ ws/gl\ d\ ws$. (*Remember*: The order in which the genes are listed is not necessarily the correct order on the chromosome.) Sample data on phenotypes of progeny from such a cross is shown in Table 6.12.

The most numerous classes are parental types, indicating linkage of genes on a single chromosome. The remaining classes must be the result of recombination. Before counting the average number of crossovers, notice that we can now observe evidence of either one or *two* crossovers. Because we are mapping three genes, a first crossover could occur between the leftmost gene and the central gene and a second crossover between the central gene and the rightmost gene. We will use the terminology *single crossover* when only one of these is observed and *double crossover* when both are observed.

► From Table 6.12, what are the likely phenotypes of double crossovers? Of single crossovers?

Notice that two of the phenotypic classes are extremely rare and four of the classes are of intermediate size. Because a double crossover is much less likely than a single crossover, this identifies the phenotypic classes that correspond

to double crossovers. Actually, we'll be able to figure out the gene order too now, if we examine the genotype of individual chromosomes carefully.

- If the genes are arranged along the chromosome in order *gl d ws*, what gametes would be produced from a double crossover in the heterozygous parent? What if they were arranged in the order *gl ws d* or *d gl ws*?

In this testcross, crossing over only effects the gametes formed by one of the parental strains. The parental-type gametes from this line are $gl^+d^+ws^+$ and *gl d ws*, with the alleles either all wildtype or all recessive. But the phenotypic classes of the double crossovers show what the chromosome inherited from this parent must have been. The class normal leaves/dwarf/white sheaf must have arisen from gametes $gl^+d\ ws$, and the class glossy leaves/tall/normal sheaf from the complementary $gl\ d^+ws^+$.

Because the outcome of a double crossover is to exchange the middle allele in the parental types, the only way a double crossover could produce the gametes here is if the genes are ordered as *d gl ws* or *ws gl d*. The *gl* gene must be in the middle. Figure 6.6 illustrates one possible configuration for a *three-strand double crossover* in which the recombinant $d^+gl\ ws^+$ is formed.

Now we are ready to estimate genetic distances. We start by finding the distance between *d* and *gl*. Four phenotypic classes result from crossovers between *d* and *gl*: tall/glossy leaves/white sheaf ($d^+gl\ ws$) and dwarf/normal leaves/normal sheaf ($d\ gl^+ws^+$) from single crossovers, tall/glossy leaves/normal sheaf ($d^+gl\ ws^+$) and dwarf/normal leaves/white sheaf ($d\ gl^+ws$) from double crossovers. Thus, the recombination frequency

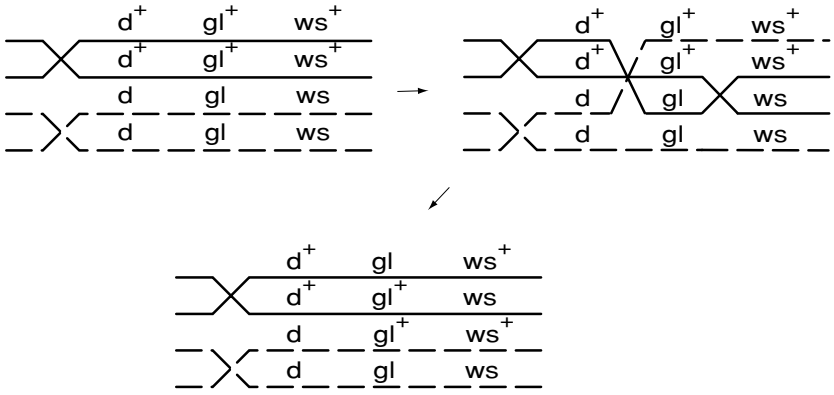

Figure 6.6. A three-strand double crossover.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

between d and gl is

$$\frac{17+15+2+1}{925} = \frac{35}{925} \approx .04.$$

Because this is small, we estimate the genetic distance as 4 cM.

Similarly, single crossovers between gl and ws produce phenotypes dwarf/glossy leaves/normal sheaf ($d\ gl\ ws^+$) and tall/normal leaves/white sheaf (d^+gl^+ws). We include the double crossover phenotypic classes in our tally, too, because a crossover occurred between the genes in them also. Thus, the recombination frequency between gl and ws is

$$\frac{154+146+2+1}{925} = \frac{303}{925} \approx .33 \text{ or } 33\ cM.$$

Thus, we estimate the genetic distance as 33 cM, but since 33 is not so small, we may worry that this estimate is not so accurate.

Note that an estimation of the distance between d and ws requires that we count *all* crossovers between the genes, so double crossovers must count as two:

$$\frac{17+15+154+146+2(2)+2(1)}{925} = \frac{338}{925} \approx .37 \text{ or } 37\ cM.$$

In particular, our estimates of genetic distance are additive, since 4 cM + 33 cM = 37 cM.

Finally, we put this together and draw the genetic map of Figure 6.7.

Return for a moment to considering only two genes, a and b. If we suspect that a and b are linked, then we might breed $a^+b^+/a\ b$ as F_1, perform a testcross with $a\ b/a\ b$, and calculate the recombination frequency. This frequency is our estimate of the genetic distance between the genes.

Notice, however, that even if a and b are located on different chromosomes, a recombination frequency can still be computed. If we did not realize they were on different chromosomes, we would count a^+b and $a\ b^+$ as single crossovers. However, because the two genes assort independently, the heterozygous parental strain produces four types of gametes, a^+b^+, a^+b, $a\ b^+$, and $a\ b$, in equal proportions. Thus, half the offspring in F_2 will show recombinant genotypes and the recombination frequency will be .5. This means we would estimate that genes on different chromosomes are 50 cM apart!

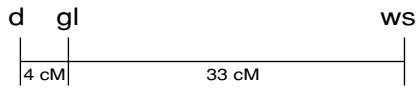

Figure 6.7. A three-gene genetic map.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

The error we have made is in assuming

$$\text{genetic distance} \approx \text{recombination frequency},$$

despite the fact that this approximation is only valid when the recombination frequency is small. Even for genes on the same chromosome, as recombination frequencies approach .5, the true genetic distance gets larger and larger. The approximation assumes multiple crossovers are rare, and that is only justifiable if the recombination frequency is small.

In genetic mapping, we must map genes that are close together first, and then build our map out from them. For example, if we want to find the distances in a chromosome with genes ordered $a - b - c - d$, it is better to calculate distances between a and b, b and c, c and d, than to try to use linkage information about only a and d. A reasonable rule of thumb is that recombination frequency is a good estimator of genetic distance when it is less than .25. Genes at a distance of $50\,cM$ or greater will assort approximately independently, as if they were on different chromosomes.

Performing testcrosses for genetic mapping of humans is of course neither ethical nor practical. Nonetheless, through pedigree analysis and somatic-cell hybridization techniques, genetic maps of the longest human chromosome have been built, with total length about 293 cM.

In addition to genetic maps, there are several other types of maps of chromosomes. A *physical map* shows markers, which might be genes or other distinguishable features, along the chromosome. Because crossover frequency does not correlate well with physical distance, such a map can look quite different, despite showing the same linear ordering of genes. *Sequencing* a chromosome to display the full structure of the DNA in terms of its constituent bases produces the highest resolution map, though genes and other features must be identified in a sequence to relate it to a genetic or physical map. Despite rapid advances in sequencing, genetic maps of the sort discussed here will remain important because of their direct applicability to problems of inheritance.

Problems

6.3.1. Three of Queen Victoria's nine children by Albert are known to have carried the X-linked allele for hemophilia. (Two of her four sons were hemophiliacs, and one of her five daughters had a hemophiliac son.) Neither she nor Albert were hemophiliacs.

a. What must have been the genotypes of Victoria and Albert? Explain how you can rule out all other possibilities.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

b. What was the probability of a son of Victoria and Albert having hemophilia? Of a daughter having hemophilia? Of a daughter being heterozygous for the allele?
c. What was the probability that exactly three of Victoria and Albert's children would carry the mutant allele?

6.3.2. Using the model Morgan developed for the X-linked, white-eyed mutant allele, compute the phenotypic ratios you would expect in the outcome of the experiment described by the data in the middle column of Table 6.9. Is the model in rough agreement with the data?

6.3.3. In another experiment, Morgan crossed white-eyed females to red-eyed males.
a. What must the genotypes of these flies be?
b. What genotypes and phenotypes would be in F_1? In what proportions?
c. If males from F_1 were crossed with females from F_1, what would be the resulting genotypes and phenotypes? In what proportion?

6.3.4. Perform a χ^2-test with $\alpha = .05$ to see if the observed data from the $X^t X^w \times X^w Y$ in the last column of Table 6.9 is consistent with expected numbers from such a cross. (Apparently, Morgan did not perform such a test.)

6.3.5. Suppose a rare disease is caused by a recessive X-linked gene, and phenotypically normal parents have a son who develops this disease.
a. If another son is born into the family, what is the probability he will develop the disease?
b. If a daughter is born into the family, what is the probability she will be a heterozygous carrier?
c. If there are two daughters in the family, what is the probability both will be carriers of the mutant allele?

6.3.6. A man with X-linked color blindness marries a woman with no history of color blindness in her family. Their daughter then marries a man with no history of color blindness and has children. What is the probability that
a. a son in the last generation will be color blind?
b. a daughter in the last generation will be color blind?
c. exactly two of three sons in the last generation will be color blind?

6.3.7. A certain allele is known to be X-linked. Determine, to the extent possible, genotypes of the parents and whether the allele is dominant or recessive if the allele is expressed in the progeny by:

a. none of the females; all of the males
b. 50% of the females; 50% of the males
c. all of the females; none of the males
d. none of the females; 50% of the males
e. 25% of the progeny

6.3.8. In a breeding experiment with flies, a particular cross produces 105 females with mutant phenotype, 98 wildtype females, and 179 wildtype males. Give a possible explanation for this outcome.

6.3.9. Vermilion eye color in *Drosophila* is caused by a recessive X-linked gene. Black body color is caused by a recessive allele on an autosome. Wildtype individuals for these genes have brick red eyes and gray body color. What phenotypic ratios are expected from the crosses:
a. gray females with brick red eyes heterozygous for both genes $\times$ black males with vermilion eyes?
b. heterozygous gray females with vermilion eyes $\times$ homozygous gray males with brick red eyes?

6.3.10. Under a hypothesis of independent assortment of genes, the cross resulting in the data shown in Table 6.11 would be expected to produce a 1:1:1:1 phenotypic ratio. Apply the χ^2-test with $\alpha = .05$ to the data to test whether the data supports rejecting a hypothesis of independent assortment.

6.3.11. Suppose a diploid organism has seven pairs of chromosomes, and each chromosome has an equal number of genes on it.
a. What is the probability that two genes chosen at random lie on distinct pairs of chromosomes?
b. Would the probability that two randomly chosen genes assort independently be greater or less than this number?

6.3.12. Suppose in the three-point test cross described by Table 6.12 you attempt to compute the genetic distance between the d and ws genes by first collapsing the table to only show information about these phenotypes.
a. Create a table like Table 6.12, but with only 4 phenotypes: tall, normal sheaf; tall, white sheaf; dwarf, normal sheaf; dwarf, white sheaf. Fill in numbers by adding appropriate entries of Table 6.12.
b. Use your table to estimate the genetic distance.
c. Why does this not agree with the estimate in the text? What is incorrect about this approach?

6.3.13. Two recessive alleles, *su* for sugary kernels and *gl* for glossy leafs, are known to exist in certain corn plants. A testcross $su^+gl^+/su\ gl \times su\ gl/su\ gl$ is performed to test for linkage. The progeny are: 198 wildtype, 228 sugary/glossy, 39 sugary, and 35 glossy. Is there evidence for linkage? If so, what is the recombination frequency between the loci for *su* and *gl*?

6.3.14. In *Drosophila*, the genes with recessive alleles *sn* for singed bristles and *m* for miniature wings are located approximately 15 *cM* apart.
   a. What sort of gametes, and in proportions, can be formed by $sn^+m^+/sn\ m$?
   b. If *Drosophila* with genotype $sn^+m^+/sn\ m$ are intercrossed, what phenotypes, and in what proportion, are the progeny?

6.3.15. Suppose two genes with alleles *a* and *b* are located 10 *cM* apart. On a different autosome, two other genes with alleles *c* and *d* are located 14 *cM* apart. Suppose individuals with genotype $a^+\ b^+/a\ b$, $c^+d^+/c\ d$ are crossed with individuals, homozygous recessive for each of these genes. What phenotypes, and in what proportions, are represented in the progeny?

6.3.16. Suppose two genes with alleles, *a* and *b*, are linked. In a heterozygote, there are two possible configurations for the chromosomes. If the genotype is $a^+b^+/a\ b$, the arrangement is called a *coupling* or *cis* configuration. If the genotype is $a^+b/a\ b^+$, the layout is known as a *repulsion* or *trans* configuration. Is it possible to use a *trans* configuration in genetic mapping? Why or why not?

6.3.17. Experimental evidence indicates that crossing over seems to be less likely near the ends or the centromere of a chromosome. Suppose two genes, *a* and *b*, are located near the centromere of a chromosome, about 5 *cM* apart. Two other genes, *c* and *d*, are located about 5 *cM* apart and about 40 *cM* away from the centromere. Which physical distance, that separating *a* and *b*, or *c* and *d*, is likely to be greater? Explain.

6.3.18. For two genes on a chromosome, give an example of a tetrad crossover configuration that results in recombinant gametes only. Why can't any tetrad configuration produce one parental-type and three recombinant gametes?

6.3.19. Suppose in a certain plant species, three genes are known to be linked. The recessive alleles for these genes are *a* for amethyst flowers, *b* for brown stalks, and *c* for curved leaves. Plants are bred with genotype

$(a^+b^+c)/(a\ b\ c^+)$, where the parentheses indicate that the order of the genes is unknown. In a testcross with $(a\ b\ c)/(a\ b\ c)$, two phenotypic classes occur in much smaller numbers: wildtype and plants with amethyst flowers, brown stalks, and curved leaves. What is the correct gene order?

6.3.20. In *Drosophila*, the genes with recessive alleles cl for clot eyes, dp for dumpy wings, and rd for reduced bristles are known to be linked.
   a. Give two different examples of appropriate testcrosses to determine the order of these genes.
   b. Suppose the phenotype wildtype eyes, dumpy wings, and reduced bristles corresponds to a recombinant from a double crossover, where the heterozygous parent had genes in the $(cl^+dp^+rd^+)/(cl\ dp\ rd)$ configuration. What is the correct gene order?

6.3.21. Suppose in a certain species three genes are linked, with alleles e for enlarged eyes, h for hairy legs, and p for prickly antennae. The wildtypes for these genes are normal eyes, hairless legs, and smooth antennae. Suppose $e\ h\ p$ is the correct gene order with e and h are located 12 cM apart and h and p are located 15 cM apart. In an experiment, $e^+h\ p/e\ h^+p^+$ individuals are testcrossed with triply homozygous recessive individuals. What are the phenotypes of the offspring and in what frequencies should these phenotypes occur?

6.3.22. For X-linked genes, you can also analyze three-point testcrosses.

In *Drosophila*, the alleles for cut wings ct, sable body s, and vermilion eyes v all determine recessive traits that are X-linked. The wildtype traits are long wings, gray body, and red eyes. Table 6.13 gives the results of a testcross of $(ct^+s^+v^+)/(ct\ s\ v)$ females with $(ct\ s\ v)$ males. Parentheses here denote unknown gene order.

Table 6.13. *Progeny of* $(ct^+s^+v^+)/(ct\ s\ v) \times (ct\ s\ v)$

Phenotype	No.
Long wings, gray body, red eyes	723
Long wings, gray body, vermilion eyes	8
Long wings, sable body, red eyes	71
Long wings, sable body, vermilion eyes	125
Cut wings, gray body, red eyes	105
Cut wings, gray body, vermilion eyes	106
Cut wings, sable body, red eyes	5
Cut wings, sable body, vermilion eyes	776

a. Notice that no data are presented on the sex of the progeny, despite the fact that X-linked genes are being investigated. Explain why it is not necessary to give that information.
b. Does the data give evidence for linkage between the three genes? Explain.
c. Determine the order of the three loci ct, s, and v and estimate the distances between them on a linkage map.

6.3.23. The occurrence of one crossover on a chromosome can inhibit the likelihood of a second crossover occurring nearby. This phenomenon, *interference*, typically takes places at distances less than 20 cM.

On chromosome III in *Drosophila*, the genes *cu* for curled wings, *Sb* for stubble bristles, and *e* for ebony body are located at 50.0 cM, 58.2 cM, and 70.7 cM, respectively. Suppose 2,000 fruit-fly progeny result from a three-point testcross.

a. Assuming only single crossovers can occur between consecutive pairs of these genes, and that there is no interference, what is the expected number of double crossovers between *cu* and *e*?
b. If interference occurs, then the observed number of double crossovers is less than expected. Define the *coefficient of coincidence, c*, to be the ratio *observed number of double crossovers: expected number of double crossovers.* If only three double crossovers are observed, what is the coefficient of coincidence?
c. The level of interference is measured by $I = 1 - c$. Explain why I is a reasonable way to quantify interference. If there is no interference, what is the value of I?
d. What is the level of interference in the testcross above?

Projects

1. Use the outcomes of simulated experiments to map genes.

The MATLAB program `genemap` will perform simulated 2- and 3-point crosses for 6 autosomal genes in *Drosophila*. (It is easily modified to simulate data for mouse genes as well.) Perform a number of such crosses and construct a genetic map from your results.

Suggestions

- Pick a reasonable number of progeny to produce, keeping in mind the laboratory and time resources necessary for real experiments.
- Record all the data from each of your crosses, to present it as support for your map.
- These genes may or may not all be on the same chromosome.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

- Because in a laboratory experiment, each cross could require much time and labor, try to keep the number of crosses you do relatively small while still gathering sufficient data. Also, 3-point crosses should be viewed as more work than 2-point ones, since they would require more breeding to prepare the lines.
- Once you have produced a genetic map, use it to predict the outcome of some crossing experiments you did not do previously. Then perform the experiments. Are the results consistent with your map? Explain any discrepancies.
- If you repeat your work using crosses that produce 10 times as many progeny, how does that affect your map? Of which map would you be more confident?
- Can you back up a claim that several genes are on different chromosomes with evidence? Can you back up a claim that several genes are on the same chromosomes with evidence?

6.4. Gene Frequency in Populations

So far, we have focused on one parental cross at a time in our models of genetics. As valuable as this may be for basic biological understanding, and for medical applications, it has neglected the larger picture. In evolution, the genetic make-up of species and populations may change over time. Some traits may be lost, other new ones arise, while some persist unchanged. Though chance plays a large role in the inheritance of traits in a single parental cross, understanding how this plays out in the evolution of a population requires mathematical modeling.

Suppose several alleles of a gene are present in a population. You might imagine a gene that determines eye color, or one that can affect the fertility of its carrier. Does the proportion of each allele change over time, or does it remain fixed? The answer might depend on the particular gene, of course, since an allele decreasing fertility seems more likely to disappear from a population than one that affects a more superficial trait such as eye color. Nonetheless, alleles that seem innocuous are observed to disappear from certain breeding populations.

We'll first study a type of equilibrium of genetic composition of a population and then investigate models of two forces tending to change the composition.

Let's focus on a single gene in a large population. To describe the variability of this gene among the population members, we use *allele frequencies*. Although technically these are relative frequencies, or the proportions of all

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

alleles that are of a certain type, we will use the simpler term "frequency" throughout this section.

The MN blood typing system in humans provides a good example of how we can estimate allele frequencies. The presence of each of the alleles M and N can be detected through antigen tests. A person with genotype MM has type M blood, and a person with genotype NN has type N blood. A heterozygote MN will test positive for both alleles and so has type MN blood. (The two alleles M and N are thus codominant, as both are equally expressed in the phenotype.)

Suppose, in a certain population, that 60 individuals have type M blood, 101 individuals type MN blood, and 53 individuals type N blood, for a total population size of 214. Because each person carries two alleles of the gene, there are a total of $2(214) = 428$ alleles in this data. To determine the frequency of M alleles, we note that each person of M blood type carries 2, those of type MN carry 1, and those of type N carry 0. Thus, the frequency of the alleles is

$$M : \frac{2(60) + 1(101)}{428} \approx .52, \quad N : \frac{1(101) + 2(53)}{428} \approx .48,$$

and of course these add to give 1.

Notice the genotype frequencies in this population are

$$MM : \frac{60}{214} \approx .28, \quad MN : \frac{101}{214} \approx .47, \quad NN : \frac{53}{214} \approx .25.$$

We can use these to calculate allele frequencies also, but because each genotype involves 2 alleles, we have to divide by 2 to account for the change in the number of objects:

$$M : \frac{2(.28) + 1(.47)}{2} = .28 + \frac{1}{2}(.47) \approx .52,$$

with a similar calculation giving the frequency of N.

Random mating and Hardy-Weinberg equilibrium. Suppose now we have a large population with the allele frequencies of the M and N blood types as calculated above. As new generations are produced, do these frequencies change?

To explore what might happen in future generations, we have to make some assumptions about the mating process. The simplest model, *random mating*, is that the genotypes of offspring are determined by the random pairing of all gametes that might be produced from current organisms. This means a given

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

gamete is equally likely to unite with any other gamete. Because our model does not track the sex of the source of the gamete, this is of course impossible for many organisms. However, assuming allele frequencies are the same in the two sexes makes the model more reasonable.

Under the random mating model, the probability of various genotypes occurring in the next generation can be calculated simply; just multiply the appropriate allele frequencies. Since picking two gametes to unite can be viewed as two independent events, the multiplication rule of probability applies. For instance, using the previously described blood type allele frequencies, the probability that an arbitrary zygote has genotype MM is $(.52)(.52) = .2704$, because .52 is the probability of picking a gamete with the M allele. Similarly, the expected frequency of the NN genotype in the offspring is $(.48)(.48) = .2304$. Since the MN genotype can be formed in two ways, MN or NM, we find, by the addition rule for disjoint probabilities, that expected frequency is $2(.52)(.48) = .4992$.

Notice that we could have used binomial probabilities in calculating the genotype frequencies instead. For instance, if we define a success as having an M allele, then $p = \mathcal{P}(S) = .52$, $q = .48$, the number of trials is $n = 2$, and the frequency of the MN genotype is

$$\mathcal{P}(\text{one success in two trials}) = \binom{2}{1}(.52)(.48).$$

► Compare the genotype frequencies of the new generation, .2704, .4992, and .2304, with the original. Did they change? How?

Now, let's calculate the allele frequencies in the new generation:

$$M : .2704 + \frac{1}{2}(.4992) = .52, \quad N : \frac{1}{2}(.4992) + .2304 = .48.$$

Remarkably, these allele frequencies are exactly the same as the original ones. Although the genotype frequencies changed a bit, the allele frequencies did not change in the new generation.

► Repeating these calculations for a third generation would produce exactly the same allele frequencies *and* genotype frequencies as in the second generation. Explain why.

Under random mating, then, we have found that the allele frequencies are in a state of equilibrium. This equilibrium is called the *Hardy-Weinberg equilibrium*, after the British mathematician Hardy and the German physician Weinberg who independently discovered it.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Let's work more theoretically with the allele frequencies to see why such an equilibrium state exists. We continue to focus our attention on a diallelic gene, with alleles a^+ and a. Let p denote the frequency of a^+ in the population and q the frequency of a, so $p + q = 1$. The assumption of random mating is what allows us to calculate the frequency of a^+ in the next generation: each allele in a second generation individual is a^+ with probability p, or a with probability q.

In the next generation, then, the allele a^+ occurs in genotypes a^+a^+ and a^+a, which have frequencies, p^2 and $2pq$, respectively. However, only half of the alleles in a^+a genotypes are wildtype. Thus, the frequency of the allele a^+ in the progeny equals

$$p^2 + \left(\frac{1}{2}\right) 2pq = p^2 + pq = p(p+q) = p(1) = p,$$

and the allele frequencies are constant from generation to generation.

- Whether the assumption of random mating is reasonable for humans might depend on what gene is being considered. Give examples of some traits for which you think it is reasonable and some for which it might not be.
- If a population is in Hardy-Weinberg equilibrium, what sorts of things not included in our model might move it away from equilibrium?

You might have noticed in the MN blood typing examples previously described that codominance allowed us to detect heterozygotes in the population and then to compute both genotype and allele frequencies. If a gene has a completely dominant allele, however, it may be difficult to distinguish between homozygous dominant and heterozygous individuals. Nonetheless, if we assume the population is in Hardy-Weinberg equilibrium, we can still estimate allele and genotype frequencies.

For example, in the United States, approximately 1 in every 3,700 individuals suffers from cystic fibrosis, the most frequent serious genetic disease of childhood, causing severe respiratory and digestive problems. Because cystic fibrosis is caused by a recessive autosomal allele, we estimate the frequency of homozygous recessives is $1/3700$. Thus, we estimate

$$q^2 \approx \frac{1}{3700}, \quad \text{so } q \approx \frac{1}{\sqrt{3700}} \approx .0164 \quad \text{and}$$
$$p = 1 - q \approx 1 - .0164 = .9836.$$

With these values, an estimate for the proportion of heterozygotes in the

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

population is $2pq = 2(.9836)(.0164) \approx .0323$. In other words, roughly 3% of the population carries the mutant allele without showing signs of the disease.

In nature, many alleles are not in Hardy-Weinberg equilibrium. In fact, evolution occurs through changing allelic frequency; so, if all genes were in equilibrium, there could be no evolution. Indeed, many real-life circumstances lead to non-equilibrium situations: In a certain population, mating might fail to be random with particular phenotypes preferring to mate with similar phenotypes (assortative mating), or individuals might migrate into or out of a subpopulation, disrupting an equilibrium. Differences in viability or fertility may result in certain genotypes having a higher survival rate and being more likely to reproduce. Spontaneous mutations may introduce new alleles into a population, changing allele frequencies. Even the size of a population may alter allele frequencies, because random forces may influence the genetic makeup in small populations. Although a Hardy-Weinberg equilibrium is appealing mathematically, it is not a long-term feature of the natural world.

Fitness and selection. Mutation and natural selection, two potent forces of evolutionary change, bring about changes in allele and genotype frequencies. Mutations produce new alleles, and organisms with a new genotype may have a changed ability to survive and reproduce. Only the genes of organisms that successfully produce offspring appear in future population members. Genes of organisms that are less well adapted to their environment may be passed along to the next generation in smaller numbers. Thus, the gene pool may be in constant flux as mutations introduce variability that selection may then weed out.

Geneticists use the term *fitness* for a measure of the ability of an organism to survive and reproduce. Suppose, for two alleles of a gene, A and a, an individual with genotype AA is the most fit. Then, we will define its *relative fitness*, w_{AA}, to be 1, and assign fitness values w_{Aa} and w_{aa} between 0 and 1 to the other two genotypes. For example, if relative fitness values are given by

$$w_{AA} = 1,\ w_{Aa} = .98,\ w_{aa} = .92,$$

then in this species the most fit genotype is AA, and heterozygotes are more fit than aa homozygotes.

- With these fitness values, do you think the allele frequency of A will increase or decrease over time?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Of course, there are many other possible relationships between relative fitness values. If A is completely dominant over a, and fitness depends on phenotype, then $w_{AA} = w_{Aa}$. If the homozygous recessive genotype is more fit, then we have $w_{aa} = 1$ and $0 \leq w_{AA} = w_{Aa} < 1$. In the exercises, some of the many other cases will be investigated.

Although relative fitness can describe selective advantage, sometimes alternate terminology is used, focusing on the selective disadvantage of a genotype. A genotype with relative fitness w is said to have *selection coefficient* $s = 1 - w$. In our previous example, the selection coefficients are 0, .02, and .08, respectively. With a selection coefficient of .08, we see that the homozygous recessive genotype is the genotype whose members will pass on the fewest genes to progeny.

We can now model how allele frequencies change because of selection. Suppose that A occurs with frequency p in the population, so a occurs with frequency $q = 1 - p$. Our model will track how p changes with time, under the assumption that mating is random.

At fertilization, gametes randomly unite to produce genotypes AA, Aa, and aa, in proportions

$$p^2,\ 2pq,\ q^2.$$

The relative fitness values then account for the competition in survival and reproduction between the genotypes as these zygotes mature and produce new gametes. Thus, the measures of the contribution of each of these genotypes to the next collection of gametes are the products

$$w_{AA}p^2,\ w_{Aa}2pq,\ w_{aa}q^2.$$

Now, because the relative fitness coefficients are less than or equal to 1, we see

$$w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2 \leq p^2 + 2pq + q^2 = (p+q)^2 = 1.$$

Therefore, we must renormalize (i.e., divide through by the quantity $w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2$) to calculate the successful contribution of gametes to the genotype proportions of the next generation, obtaining

$$\frac{w_{AA}p^2}{w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2},\ \frac{w_{Aa}2pq}{w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2},$$

$$\frac{w_{aa}q^2}{w_{AA}p^2 + w_{Aa}2pq + w_{aa}q^2}.$$

Finally, because all the alleles contributed by the AA genotype are A, but

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

only half the alleles contributed by the Aa genotype are, we find

$$p_{t+1} = \frac{w_{AA}p_t^2}{w_{AA}p_t^2 + w_{Aa}2p_tq_t + w_{aa}q_t^2} + \frac{1}{2}\frac{w_{Aa}2p_tq_t}{w_{AA}p_t^2 + w_{Aa}2p_tq_t + w_{aa}q_t^2}$$
$$= \frac{w_{AA}p_t^2 + w_{Aa}p_tq_t}{w_{AA}p_t^2 + w_{Aa}2p_tq_t + w_{aa}q_t^2}$$

► Express this in terms of p_t alone, with no q_t.

Let's consider a concrete example. Suppose, initially, 70% of the alleles are A. Thus, $p_0 = .7$ and $q_0 = .3$. If all genotypes are equally fit, then $w_{AA} = w_{Aa} = w_{aa} = 1$, and we find

$$p_1 = \frac{p_0^2 + p_0q_0}{p_0^2 + 2p_0q_0 + q_0^2} = \frac{.49 + .21}{1} = .7,$$

which illustrates the Hardy-Weinberg equilibrium. If, however, relative fitness values $w_{AA} = 1$, $w_{Aa} = .98$, and $w_{aa} = .92$ describe the genotypes, then

$$p_1 = \frac{p_0^2 + (.98)p_0q_0}{p_0^2 + (.98)2p_0q_0 + (.92)q_0^2} = \frac{.49 + (.98).21}{.9844} = .7068.$$

As you might expect, the allele frequency of A has increased slightly, from .7 to .7068, at the expense of the allele a.

Iterating the model over a few generations produces Figure 6.8. Since the genotypes are increasingly fit according to the presence of the allele A, over many generations A becomes fixed in the population and the recessive allele dies out.

This model becomes even more interesting for parameter choices where the outcome is less intuitive. What might happen if a recessive allele was the most fit? Would it be fixed eventually, or would the fact that it was only expressed in homozygotes give it too weak an influence to eventually predominate? Or, what if the heterozygotes were the most fit genotype? The outcome of such a situation is hard to predict without a mathematical model. These questions are not simply a result of mathematical curiosity, as a few biological examples show:

- In a certain species of moths, a dominant allele is associated with dark coloring. Homozygous recessives are light-colored. If a moth population lives in a forest with dark-colored trees, the light-colored moths are at a competitive disadvantage, as their predators can more easily see them. If

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

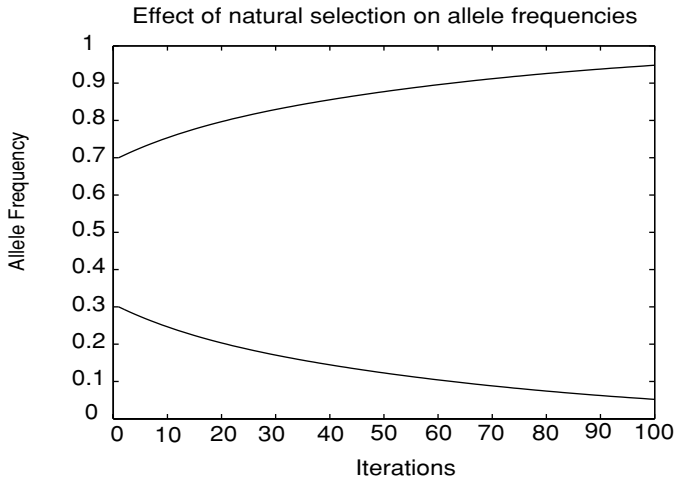

Figure 6.8. Allele frequencies of A (top) and a (bottom); relative fitness values $w_{AA} = 1$, $w_{Aa} = .98$, and $w_{aa} = .92$.

the tree bark tends to be lighter-colored, then light-colored moths are more likely to survive.

- In humans, the often-fatal disease sickle-cell anemia is associated with a homozygous recessive genotype. In certain parts of the world, the recessive allele is quite common – by some estimates about as high as 19%. Researchers have discovered that heterozygotes have an increased resistance to malaria, and thus a greater fitness in a tropical climate.

In the exercises, we will explore a number of scenarios for the effects of natural selection:

Selection for A: favors the dominant allele and associated phenotypes.
Selection against A: favors homozygous recessives.
Heterozygote Advantage or Overdominance: favors heterozygotes at the expense of homozygotes.
Homozygote Advantage: favors homozygotes, at the expense of heterozygotes.

The frequency of an allele may rise or fall, depending on the forces of selection.

Genetic drift. So far, our models addressing allele frequencies have tacitly assumed that the population under study was large. For instance, we assumed we were modeling a large population when we argued that because a certain

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

Table 6.14. *Probabilities That Exactly k of 4 Alleles Are A*

k	0	1	2	3	4
$\mathcal{P}(k)$	.0625	.25	.375	.25	.0625

proportion of the gametes had an allele, then the same proportion of the gametes that successfully united would have that allele. Even if half the gametes have an allele A, if we randomly pick gametes to unite, we might pick more or less than half As to form the next generation. In a small population, any deviation from half might be proportionally large, and thus proportionally greater than you are likely to have in a large population. In other words, small populations are more greatly affected by chance than are large ones.

For a concrete illustration of this, imagine a very small population of 2 individuals of genotypes Aa and Aa. Then, the alleles A and a appear in the gamete pool in proportions .5 and .5, and so random mating implies that each offspring will have genotype AA (or aa) with probability .25, and genotype Aa with probability .5.

However, if the new generation also has size 2, then to determine the alleles in this generation, we simply pick four specific gametes out of the pool. Using the binomial distribution, the probability of having exactly two of each allele in the next generation is

$$\binom{4}{2}(.5)^2(.5)^2 \approx .375.$$

This means that the probability that the allele frequencies remain stable is only 37.5%, and the more likely scenario is that allele frequencies will change. Furthermore, any change in the allele frequency must be at least .25, because there are only four alleles total in this small population. Thus, a reasonably large change is quite likely.

It might seem that this result contradicts the ideas underlying the Hardy-Weinberg equilibrium for allele frequencies. However, calculating the probabilities that exactly k of 4 alleles are A for $k = 0, 1, 2, 3$, and 4 as in Table 6.14, we see the most likely outcome is that the allele frequencies represented in the two offspring will be $p = q = .5$, the same frequencies of the parental generation and just as Hardy-Weinberg predicts. However, this most likely outcome is not very likely.

If a population is large – say 3,000 heterozygotes producing 3,000 offspring – then producing a table like Table 6.14 also shows that some change in allele numbers is likely. However, the likely size of this change is much

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

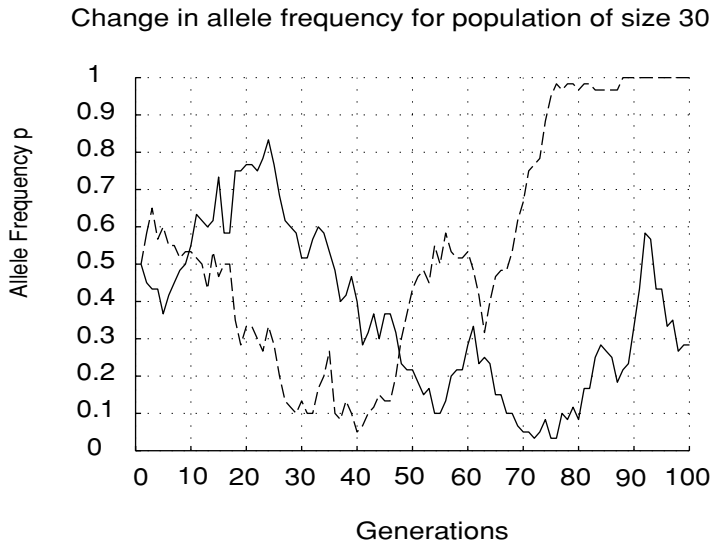

Figure 6.9. Two examples of genetic drift; population size $N = 30$.

smaller proportionally than for the two individual case. Rather than changes in allele frequencies of magnitude .25, tiny changes typically occur. Thus, the Hardy-Weinberg values are a more accurate estimate of what actually happens.

For large populations, we lose little by ignoring chance fluctuations. If a population is small, then chance fluctuations are much more important and may in fact predominate. The phenomenon of chance changes in allele frequencies dominating other factors in small populations is known as *genetic drift*.

Genetic drift may be modeled by fixing a population size N and initial allele frequencies. Then, a new generation of alleles is chosen according to the probabilities calculated by the binomial distribution. Using the new allele frequencies, this process is repeated for the next generation, and so on. Because of the random choices made at each generation, no two simulations are likely to be identical.

Figure 6.9 shows two simulations of allele frequency p over a number of generations. In both plots, the population is small, $N = 30$, and the initial value is $p = .5$. Notice the random fluctuation of the frequency p, and that whether the allele remains fixed in the population or is removed entirely is a matter of chance.

Using only concepts introduced here, it is easy to imagine a more sophisticated model that combines genetic drift with selection. But models of genes

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

with more alleles, or of several genes that collectively determine traits affecting fitness, are also possible. Modeling the creation of new alleles through mutation, along with their possible elimination or fixation through selection, also leads to interesting insights. We have really only scratched the surface of mathematical models in population genetics.

Problems

6.4.1. An autosomal recessive allele *ct* causes curly tails in mice. Suppose, in a certain population of 450 mice, 441 mice have normal tails and 9 have curly tails, and that the allele frequencies are in Hardy-Weinberg equilibrium.
   a. Estimate the allele frequency of *ct*.
   b. What percentage of the mice population is heterozygous for this gene?

6.4.2. Color blindness is an X-linked trait that occurs in about 8% of human males.
   a. Give the allele frequencies for this gene. (Assume the frequencies p and q are the same in both genders, and are in equilibrium.)
   b. Approximately what percentage of the female population is color blind? What percentage of the female population with normal vision carries the mutant allele?

6.4.3. Suppose a randomly mating population segregating two alleles is in Hardy-Weinberg equilibrium.
   a. What are the allele frequencies p and q if the frequency of heterozygotes is .4? If the frequency of heterozygotes is H?
   b. Express the frequency of heterozygotes in terms of p. What values of p and q maximize this frequency? (Either graphing or calculus can be used to answer this.)

6.4.4. There is a strong connection between certain powers of polynomials and genotype frequencies in simple situations.
   a. Expand the binomial power $(p+q)^2$ and explain the meaning of each summand in terms of genotype frequencies for a diallelic gene.
   b. If a gene has multiple alleles, *multinomial expansions* are related to genotype frequencies. Suppose a gene has 3 alleles, occurring in frequencies p, q, and r. Expand $(p+q+r)^2$ and relate each term in the expansion to genotype frequencies.
   c. Does the concept of a Hardy-Weinberg equilibrium make sense for the 3 allele situation? Explain.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

6.4.5. The genetics of the ABO blood typing system was explained in Problem 6.1.16.

a. In ABO blood-typing studies in an isolated community, 32% of the population have type A blood, 15% type B blood, 4% type AB blood, and 49% type O blood. Determine the allele frequencies I^A, I^B, and I^O in this community.

b. In the United States, approximately 40% of the population have type A blood, 11% type B blood, 5% type AB blood, and 44% type O blood. Give the system of equations that describes the blood-type frequencies in terms of the allele frequencies I^A, I^B, and I^0. Can you solve this system? If not, explain the difficulty and its biological implications.

6.4.6. Suppose a gene has 3 alleles in equilibrium in a randomly mating population. To find allele frequencies for the population, what is the minimum number of phenotype frequencies you must know? Answer the same question for n alleles.

6.4.7. Although a Hardy-Weinberg equilibrium may exist in a well-mixed population, over expansive geographic areas, natural barriers often cause variations in local equilibrium frequencies.

Suppose two lakes separated by a short distance are populated with the same species of fish and that both lakes are in an equilibrium state. In the first lake, the frequency of a particular allele a^+ is p_1. In the second lake, the frequency of a^+ is p_2. After a flood, the two lakes are merged, and one lake is formed. Suppose both lakes contained the same number N of fish.

a. What is the frequency p of the allele a^+ in the fish in the large lake after the flood?

b. What are the genotype frequencies immediately after the flood? What would a Hardy-Weinberg equilibrium predict for the genotype frequencies? Explain why these two answers do not agree.

6.4.8. Show the selection model simplifies considerably if $w_{AA} = w_{Aa} = w_{aa} = 1$. Using these relative fitness values, give the simplest formula possible for p_{t+1} in terms of p_t. Explain the relationship of your formula to Hardy-Weinberg equilibrium.

6.4.9. Investigate the behavior of the selection model experimentally, using a computer program such as `onepop`, for each set of relative fitness values below. Describe your observations on the model's behavior, including likely equilibria and their stability. Are the behaviors you see biologically reasonable?

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

a. $w_{AA} = 1$, $w_{Aa} = .98$, and $w_{aa} = .92$ (dominant advantage)
b. $w_{AA} = .92$, $w_{Aa} = .98$, and $w_{aa} = 1$ (recessive advantage)
c. $w_{AA} = 1$, $w_{Aa} = .92$, and $w_{aa} = 1$ (homozygous advantage)
d. $w_{AA} = .92$, $w_{Aa} = 1$, and $w_{aa} = .92$ (heterozygous advantage).

6.4.10. In mice, homozygotes for the yellow-lethal allele, Y^l, die in embryonic stage, while heterozygotes have yellow fur. What are reasonable values to use in the selection model for the selection coefficients for the three genotypes? Use a computer program such as `onepop` to investigate the model, and describe your results. Does the allele persist in the population?

6.4.11. Relative fitness values $w_{AA} = 0$, $w_{Aa} = w_{aa} = 1$ describe a special case of the selection model.
a. Interpret these biologically.
b. Show that with these values the model is simply

$$p_{t+1} = \frac{p_t}{1 + p_t}.$$

c. Show that the explicit formula

$$p_t = \frac{p_0}{1 + tp_0}, \quad t = 1, 2, 3, \ldots$$

gives allele frequencies for this model.

6.4.12. Relative fitness values $w_{AA} = w_{Aa} = 1$, $w_{aa} = 0$ describe a special case of the selection model.
a. Interpret these biologically.
b. Give the simplest formula you can expressing p_{t+1} in terms of p_t.
c. Find an explicit formula for p_t in terms of p_0 and t.

6.4.13. Find all equilibria for the selection model as follows:
a. Express the equilibrium equation that p^* must satisfy in the form of a cubic polynomial $= 0$. This shows there are at most three equilibria.
b. Two equilibria are easy to guess. (What possible allele frequencies would not change, no matter what the relative fitness values were?) What are they?
c. Use your guesses in part (b) to help you factor the cubic polynomial in part (a) completely.
d. Use part (c) to show the third equilibrium can be written as

$$\frac{(w_{aa} - w_{Aa})}{(w_{aa} - w_{Aa}) + (w_{AA} - w_{Aa})}.$$

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

6.4.14. The third equilibrium for the selection model that was found in the preceeding problem is only biologically meaningful if it is a possible value for an allele frequency.

a. Explain why the third equilibrium is only biologically meaningful if

$$(w_{aa} - w_{Aa})(w_{AA} - w_{Aa}) > 0.$$

b. Explain why the third equilibrium is only biologically meaningful if either $w_{AA} > w_{Aa}$ and $w_{aa} > w_{Aa}$ (homozygote advantage), or if $w_{AA} < w_{Aa}$ and $w_{aa} < w_{Aa}$ (heterozygote advantage).

6.4.15. Use a program such as `cobweb` to investigate the stability of the selection model equilibria under the following conditions. Use a variety of parameter choices for each. Express your conclusions in biological terminology.

a. $w_{AA} > w_{Aa}$ and $w_{aa} > w_{Aa}$ (homozygote advantage)

b. $w_{AA} < w_{Aa}$ and $w_{aa} < w_{Aa}$ (heterozygote advantage).

6.4.16. In the selection model, the quantity

$$\overline{w}_t = w_{AA}p_t^2 + w_{Aa}2p_tq_t + w_{aa}q_t^2$$

is called the *mean fitness* of the population at time t. It is possible to show that $\overline{w}_{t+1} \geq \overline{w}_t$. Why is such a result reasonable biologically?

6.4.17. Use a computer program, such as `genesim` to explore the phenomenon of genetic drift. For a population of size $N = 30$, begin with equal allele frequencies and do several simulations. Repeat for $N = 300$ and $N = 3000$. Describe your observations on how population size affects drift.

6.4.18. The program `genesim` can model genetic drift with selection effects due to varying relative fitness levels of genotypes. For a population size that exhibits strong drift when all genotypes have the same fitness, run simulations with interesting choices of relative fitness values. Describe your observations and discuss whether they seem biologically reasonable.

6.4.19. What is the expected value of the number of A alleles in the situation described by Table 6.14? How does this fit with the idea of Hardy-Weinberg equilibrium?

6.4.20. In a population of size N, if genetic drift causes changes in allele frequencies p and q, then genotype frequencies change, too. One way

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

to measure the effect of genetic drift is by monitoring the frequency H of heterozygotes, the *heterozygosity*, of a population.

a. If genetic drift tends to eliminate an allele, what will the effect be on the value of H over time? Explain.
b. A good model (which we will not justify here) to describe the effect of genetic drift on the heterozygosity of a population is $H_{t+1} = (1 - \frac{1}{2N})H_t$. Use the program `onepop` to explore the effect of population size on genetic drift and heterozygosity. Start with an initial value of $H_0 = .5$ and vary the population size N. What happens to H if $N = 100$? If $N = 1{,}000$? If N is huge? How would your answers change if the initial value was $H_0 = .2$ or $H_0 = .9$?
c. Give a formula for H_t in terms of N, H_0, and t.

Projects

1. Investigate the phenomenon of genetic drift in a simulated population.

Study a gene with two alleles, A and a, that occur in a diploid population of size N in frequencies p and q. Assume that these alleles are *selectively neutral* (i.e., the resulting genotypes are all equally fit).

Use the MATLAB program `genesim` to observe changes in allele frequencies in a simulated population over a number of generations. This program assumes that the population size N remains constant from generation to generation and that mating is random.

Explore the effect of genetic drift on allele frequencies under a variety of assumptions.

- The population size N is small, medium, or large.
- The initial allele frequency of A is $p_0 = .5$, $p_0 > .5$, or $p_0 < .5$.

The main issues to consider are:

1. What happens to the allele frequency p over the long run? Is it stable? Does the allele A become fixed in the population? Is A eliminated entirely? If either of these happens, how quickly does it occur?
2. How does the population size affect your answer to question 1 above?

Suggestions

- To get a feel for the effects of genetic drift, use the program `genesim` to explore changes in allele frequencies for lots of reasonable choices of N and p_0. Make a note of any unusual behavior and try to explain it.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

- After a large number k of generations, how does the allele frequency p_k compare with p_0? Is there a tendency for fixation or elimination of an allele? Explore this question for different population sizes.
- If $p_0 = .5$, how likely is it that A becomes fixed in the population? For fixed N, do many simulations, record the results, and from them estimate the probability of fixation. Repeat for other N.
- Investigate the last question for specific $p_0 > .5$ and $p_0 < .5$.
- Does genetic drift tend to increase or decrease genetic variation within a population? How does the population size affect your answer?
- If one population is separated into two populations by migration (early humans leaving Africa; farm-raised fish being released into two lakes), what effect might genetic drift have on the variability between the two populations?
- If you perform many genetic drift simulations for a fixed value of p_0 and N and average the values p_{50}, what will you get? Does your answer depend on the initial value p_0? N?

2. For a gene with two alleles, A and a, both the simple selection model and genetic drift often lead to fixation of one of the alleles and elimination of the other. Why, then, do we observe so many genes with multiple alleles in real populations? Are all of them either selectively neutral, or in populations so large that drift is negligible?

   Explore and discuss one or more of the following models that offer further reasons for the stability of *polymorphic* genes.

   - *Heterozygote advantage*: In this selection model, $w_{Aa} > w_{AA}$ and $w_{Aa} > w_{aa}$. (This is the mechanism by which the persistence of the sickle cell allele is generally explained.)
   - *Frequency-dependent selection*: In this type of selection model, the fitness coefficients depend on allele frequencies. One example is

   $$w_{AA} = 1 - up^2,\ w_{Aw} = 1 - u2pq,\ w_{aa} = 1 - uq^2,$$

   for some value of u between 0 and 1. In this model, the more prevalent an allele is, the less its fitness. (In certain plants, pollen with one allele can only successfully fertilize plants with other alleles, giving rare alleles an advantage.)
   - *Mutation-selection balance*: This model modifies the classical selection model to account for recurrent mutations that continually renew the stock of an allele that might otherwise disappear. For instance, if a fraction μ of alleles that would have been A in each new generation

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press

mutate to a, and p_t tracks the frequency of A, such a model is

$$p_{t+1} = \frac{w_{AA}p_t^2 + w_{Aa}p_tq_t}{w_{AA}p_t^2 + w_{Aa}2p_tq_t + w_{aa}q_t^2}(1-\mu).$$

Suggestions

- Investigate these models experimentally using `onepop` and `cobweb` for a variety of parameter choices. Describe your observations and insights.
- If possible, compute equilibria for the models and discuss their stability. (If you cannot do this in general, at least do it for a few parameter choices, or by making special choices, such as $w_{AA} = w_{Aa} = 1$, $w_{aa} = 1 - s$ in the selection-mutation model.)
- *Meiotic drive*, the preferential creation of gametes of a certain type, is another mechanism that can lead to polymorphic stability. Modify the basic selection model to take meiotic drive into account and analyze your model.

https://doi.org/10.1017/CBO9780511790911.008 Published online by Cambridge University Press