

4

PROBABILITY AND STATISTICS

STUDY OBJECTIVES

1. To understand the rules of probability and how they apply to genetics 71
2. To understand the use of the chi-square statistical test in genetics 74

STUDY OUTLINE

- Probability** 71
 - Types of Probabilities 71
 - Combining Probabilities 71
- Use of Rules** 72
- Statistics** 74
 - Hypothesis Testing 74
 - Chi-Square 76
 - Failing to Reject Hypotheses 77
- Summary* 78
- Solved Problems* 78
- Exercises and Problems* 79
- Critical Thinking Questions* 81



An agricultural worker studies variability in plants in a greenhouse. Probability influences the differences among organisms. (© David Joel/Tony Stone Images.)

In an experimental science, such as genetics, scientists make decisions about hypotheses on the basis of data gathered during experiments. Geneticists must therefore have an understanding of probability theory and statistical tests of hypotheses. Probability theory allows geneticists to construct accurate predictions of what to expect from an experiment. Statistical testing of hypotheses, particularly with the chi-square test, allows geneticists to have confidence in their interpretations of experimental data.

PROBABILITY

Part of Gregor Mendel's success was due to his ability to work with simple mathematics. He was capable of turning numbers into ratios and deducing the mechanisms of inheritance from them. Taking numbers that did not exactly fit a ratio and rounding them off to fit lay at the heart of Mendel's deductive powers. The underlying rules that make the act of "rounding to a ratio" reasonable are the rules of probability.

In the **scientific method**, scientists make predictions, perform experiments, and gather data that they then compare with their original predictions (see chapter 1). However, even if the bases for the predictions are correct, the data almost never exactly fit the predicted outcome. The problem is that we live in a world permeated by random, or **stochastic**, events. A bright new penny when flipped in the air twice in a row will not always give one head and one tail. In fact, that penny, if flipped one hundred times, could conceivably give one hundred heads. In a stochastic world, we can guess how often a coin should land heads up, but we cannot know for certain what the next toss will bring. We can guess how often a pea should be yellow from a given cross, but we cannot know with certainty what the next pod will contain. Thus, we need **probability theory** to tell us what to expect from data. This chapter closes with some thoughts on statistics, a branch of mathematics that helps us with criteria for supporting or rejecting our hypotheses.

Types of Probabilities

The **probability** (P) that an event will occur is the number of favorable cases (a) divided by the total number of possible cases (n):

$$P = a/n$$

The probability can be determined either by observation (empirical) or by the nature of the event (theoretical). For example, we observe that about one child in ten thousand is born with phenylketonuria. Therefore, the

probability that the next child born will have phenylketonuria is 1/10,000. The odds based on the geometry of an event are, for example, like the familiar toss of dice. A die (singular of dice) has six faces. When that die is tossed, there is no reason one face should land up more often than any other. Thus, the probability of any one of the faces being up (e.g., a four) is one-sixth:

$$P = a/n = 1/6$$

Similarly, the probability of drawing the seven of clubs from a deck of cards is

$$P = 1/52$$

The probability of drawing a spade from a deck of cards is

$$P = 13/52 = 1/4$$

The probability (assuming a 1:1 sex ratio, though the actual ratio is about 1.06 males per female born in the United States) of having a daughter in any given pregnancy is

$$P = 1/2$$

And the probability that an offspring from a self-fertilized dihybrid will show the dominant phenotype is

$$P = 9/16$$

From the probability formula, we can say that an event with certainty has a probability of one, and an event that is an impossibility has a probability of zero. If an event has the probability of P , all the other alternatives combined will have a probability of $Q = 1 - P$; thus $P + Q = 1$. That is, the probability of the completely dominant phenotype in the F_2 of a selfed dihybrid is 9/16. The probability of any other phenotype is 7/16, and when the two are added together, they equal 16/16, or 1.

Combining Probabilities

The basic principle of probability can be stated as follows: If one event has c possible outcomes and a second event has d possible outcomes, then there are cd possible outcomes of the two events. From this principle, we obtain three rules that concern us as geneticists.

To understand these rules of probability requires a few definitions. *Mutually exclusive events* are events in which the occurrence of one possibility excludes the occurrence of the other possibilities. In the throwing of a die, for example, only one face can land up. Thus, if it comes up a four, it precludes the possibility of any of the other faces. Similarly, a blue-eyed daughter is mutually exclusive of a brown-eyed son or any other combination of gender and eye color. *Independent events*, however, are events whose outcomes do not influence one another. For example, if two dice are thrown, the face

value of one die is not able to affect the face value of the other; they are thus independent of each other. Similarly, the gender of one child in a family is generally independent of the gender of the children who have come before or might come after. Finally, *unordered events* are events whose probability of outcome does not depend on the order in which the events occur; the probabilities combine both mutual exclusivity and independence. For example, when two dice (one red, one green) are thrown at the same time, we generally do not specify which die has which value; a seven can occur whether the green die is the four or the red die is the four. Similarly, the probability that a family of several children will have two boys and one girl is the same irrespective of their birth order. If the family has two boys and one girl, it does not matter whether the daughter is born first, second, or third. In general, probabilities differ depending on whether order is specified. With these definitions in mind, let us look at three rules of probability that affect genetics.

1. Sum Rule

When events are mutually exclusive, the **sum rule** is used: The probability that one of several mutually exclusive events will occur is the sum of the probabilities of the individual events. This is also known as the *either-or rule*. For example, what is the probability, when we throw a die, of its showing *either* a four *or* a six? According to the sum rule,

$$P = 1/6 + 1/6 = 2/6 = 1/3$$

2. Product Rule

When the occurrence of one event is independent of the occurrence of other events, the **product rule** is used: The probability that two independent events will both occur is the product of their separate probabilities. This is known as the *and rule*. For example, the probability of throwing a die two times and getting a four *and* then a six, in that order, is

$$P = 1/6 \times 1/6 = 1/36$$

3. Binomial Theorem

The **binomial theorem** is used for unordered events: The probability that some arrangement will occur in which the final order is not specified is defined by the binomial theorem. For example, what is the probability when tossing two pennies simultaneously of getting a head and a tail? We will look more closely at how to use the rules of probability to answer this question.

USE OF RULES

There are several ways to calculate the probability just asked for. To put the problem in the form for rule 3 is the quickest method, but this problem can also be solved by using a combination of rules 1 and 2. For each penny, the probability of getting a head (H) *or* a tail (T) is

$$\begin{aligned} \text{for H: } P &= 1/2 \\ \text{for T: } Q &= 1/2 \end{aligned}$$

Tossing the pennies one at a time, it is possible to get a head *and* a tail in two ways:

$$\begin{aligned} &\text{first head, then tail (HT)} \\ &\quad \text{or} \\ &\text{first tail, then head (TH)} \end{aligned}$$

Within a sequence (HT or TH), the probabilities apply to independent events. Thus, the probability for any one of the two sequences involves the product rule (rule 2):

$$1/2 \times 1/2 = 1/4 \text{ for HT or TH}$$

The two sequences (HT or TH) are mutually exclusive. Thus, the probability of getting either of the two sequences involves the sum rule (rule 1):

$$1/4 + 1/4 = 1/2$$

Thus, for unordered events, we can obtain the probability by combining rules 1 and 2. The binomial theorem (rule 3) provides the shorthand method.

To use rule 3, we must state the theorem as follows: If the probability of an event (X) is p and an alternative (Y) is q , then the probability in n trials that event X will occur s times and Y will occur t times is

$$P = \frac{n!}{s!t!} p^s q^t$$

In this equation, $s + t = n$, and $p + q = 1$. The symbol $!$, as in $n!$, is called **factorial**, as in “ n factorial,” and is the product of all integers from n down to one. For example, $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. Zero factorial equals one, as does anything to the power of zero ($0! = n^0 = 1$).

Now, what is the probability of tossing two pennies at once and getting one head and one tail? In this case, $n = 2$, s and $t = 1$, and p and $q = 1/2$. Thus,

$$P = \frac{2!}{1!1!} (1/2)^1 (1/2)^1 = 2(1/2)^2 = 1/2$$

This is, of course, our original answer. Now on to a few more genetically relevant problems. What is the probability that a family with six children will have five girls and one boy? (We assume that the probability of either a

son or a daughter equals 1/2.) Since the order is not specified, we use rule 3:

$$P = \frac{6!}{5!1!} (1/2)^5 (1/2)^1 = 6(1/2)^6 = 6/64 = 3/32$$

What would happen if we asked for a specific family order, in which four girls were born, then one boy, and then one girl? This would entail rule 2; for a sequence of six independent events:

$$P = 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 = 1/64$$

When no order is specified, the probability is six times larger than when the order is specified; the reason is that there are six ways of getting five girls and one boy, and the sequence 4-1-1 is only one of them. Rule 3 tells us that there are six ways. These are (letting B stand for boy and G for girl) as follows:

Birth Order					
1	2	3	4	5	6
B	G	G	G	G	G
G	B	G	G	G	G
G	G	B	G	G	G
G	G	G	B	G	G
G	G	G	G	B	G
G	G	G	G	G	B

Let us look at yet another problem. If two persons, heterozygous for albinism (a recessive condition), have four children, what is the probability that all four children will be normal? The answer is simply $(3/4)^4$ by rule 2. What is the probability that three will be normal and one albino? If we specify which of the four children will be albino (e.g., the fourth), then the probability is $(3/4)^3(1/4)^1 = 27/256$. If, however, we do not specify order,

$$P = \frac{4!}{3!1!} (3/4)^3 (1/4)^1 = 4(3/4)^3 (1/4)^1 = 4(27/256) = 108/256$$

This is precisely four times the ordered probability because the albino child could have been born first, second, third, or last.

The formula for rule 3 is the formula for the terms of the **binomial expansion**. That is, if $(p + q)^n$ is expanded (multiplied out), the formula $(n!/s!t!)p^s q^t$ gives the probability for any one of these terms, given that $p + q = 1$ and that $s + t = n$. Since there are $(n + 1)$ terms in the binomial, the formula gives the probability for the term numbered $(t + 1)$. Two bits of useful information come from recalling that rule 3 is in reality the

binomial expansion formula. First, if you have difficulty calculating the term, you can use **Pascal's triangle** to get the coefficients:

			1		
		1	1		
	1	2	1		
	1	3	3	1	
1	4	6	4	1	
1	5	10	10	5	1

Pascal's triangle is a triangular array made up of coefficients in the binomial expansion. It is calculated by starting any row with a 1, proceeding by adding two adjacent terms from the row above, and then ending with a 1. For example, the next row would be

$$1, (1 + 5), (5 + 10), (10 + 10), (10 + 5), (5 + 1), 1$$

or 1, 6, 15, 20, 15, 6, 1

These numbers give us the combinations for any $p^s q^t$ term. That is, in our previous example, $n = 4$; so we use the $(n + 1)$, or fifth, row of Pascal's triangle. (The second number in any row of the triangle gives the power of the expansion, or n . Here, 4 is the second number in the row.) We were interested in the case of one albino child in a family of four children, or $p^3 q^1$, where p is the probability of the normal child $(3/4)$ and q is the probability of an albino child $(1/4)$. Hence, we are interested in the $(t + 1)$ —that is, the $(1 + 1)$ —or the second term of the fifth row of Pascal's triangle, which will tell us the number of ways of getting a four-child family with one albino child. That number is 4. Thus, using Pascal's triangle, we see that the solution to the problem is

$$4(3/4)^3 (1/4)^1 = 108/256$$

This is the same as the answer we obtained the conventional way.

The second advantage from knowing that rule 3 is the binomial expansion formula is that we can now generalize to more than two outcomes. The general form for the **multinomial expansion** is $(p + q + r + \dots)^n$ and the general formula for the probability is

$$P = \frac{n!}{s!t!u! \dots} p^s q^t r^u \dots$$

where $s + t + u + \dots = n$ and $p + q + r + \dots = 1$ For example, our albino-carrying heterozygous parents may want an answer to the following question: If we have five children, what is the probability that we will have two normal sons, two normal daughters, and one albino son? (This family will have no albino daughters.) By rule 2, the probability of

a normal son = $(3/4)(1/2) = 3/8$
 a normal daughter = $(3/4)(1/2) = 3/8$
 an albino son = $(1/4)(1/2) = 1/8$
 an albino daughter = $(1/4)(1/2) = 1/8$

Thus:

$$P = \frac{5!}{2!2!1!0!} (3/8)^2 (3/8)^2 (1/8)^1 (1/8)^0$$

$$= 30(3/8)^4 (1/8)^1 = 30(3^4)/(8^5) = 2,430/32,768$$

$$= 0.074$$

STATISTICS

In one of Mendel's experiments, F_1 heterozygous pea plants, all tall, were self-fertilized. In the next generation (F_2), he recorded 787 tall offspring and 277 dwarf offspring for a ratio of 2.84:1. Mendel saw this as a 3:1 ratio, which supported his proposed rule of inheritance. In fact, is 787:277 "roundable" to a 3:1 ratio? From a brief discussion of probability, we expect some deviation from an exact 3:1 ratio (798:266), but how much of a deviation is acceptable? Would 786:278 still support Mendel's rule? Would 785:279 support it? Would 709:355 (a 2:1 ratio) or 532:532 (a 1:1 ratio)? Where do we draw the line? It is at this point that the discipline of statistics provides help.

We can never speak with certainty about stochastic events. For example, take Mendel's cross. Although a 3:1 ratio is expected on the basis of Mendel's hypothesis, chance could mean that the data yield a 1:1 ratio (532:532), yet the mechanism could be the one that Mendel suggested. In other words, we could flip an honest coin and get ten heads in a row. Conversely, Mendel could have gotten exactly a 3:1 ratio (798:266) in his F_2 generation, yet his hypothesis of segregation could have been wrong. The point is that any time we deal with probabilistic events there is some chance that the data will lead us to support a bad hypothesis or reject a good one. Statistics quantifies these chances. We cannot say with certainty that a 2.84:1 ratio represents a 3:1 ratio; we can say, however, that we have a certain degree of confidence in the ratio. Statistics helps us ascertain these **confidence limits**.

Statistics is a branch of probability theory that helps the experimental geneticist in three ways. First, part of statistics deals with **experimental design**. A bit of thought applied before an experiment may help the investigator design the experiment in the most efficient way. Although he did not know statistics, Mendel's experimental design was very good. The second way in which statistics is helpful is in summarizing data. Familiar terms such as *mean* and *standard deviation* are part of the body of descriptive statistics that takes large masses

of data and reduces them to one or two meaningful values. We examine further some of these terms and concepts in the chapter on quantitative inheritance (chapter 18).

Hypothesis Testing

The third way that statistics is valuable to geneticists is in the **testing of hypotheses**: determining whether to support or reject a hypothesis by comparing the data to the predictions of the hypothesis. This area is the most germane to our current discussion. For example, was the ratio of 787:277 really indicative of a 3:1 ratio? Since we know now that we cannot answer with an absolute yes, how can we decide to what level the data support the predicted 3:1 ratio?

Statisticians would have us proceed as follows. To begin, we need to establish how much variation to expect. We can determine this by calculating a **sampling distribution**: the frequencies with which various possible events could occur in a particular experiment. For example, if we self-fertilized a heterozygous tall plant, we would expect a 3:1 ratio of tall to dwarf plants among the progeny. (The 3:1 ratio is our hypothesis based on the assumption that height is genetically controlled by one locus with two alleles.) If we looked at the first four offspring, what is the probability we would see three tall and one dwarf plant? We can calculate the answer using the formula for the terms of the binomial expansion:

$$P = \frac{4!}{3!1!} (3/4)^3 (1/4)^1 = 108/256 = 0.42$$

Similarly, we can calculate the probability of getting all tall ($81/256 = 0.32$), two tall and two dwarf ($54/256 = 0.21$), one tall and three dwarf ($12/256 = 0.05$), and all dwarf ($1/256 = 0.004$) in this first set of four. Table 4.1 shows this distribution, as well as the distributions for samples of eight and forty progeny. Figure 4.1 shows these distributions in graph form.

As sample sizes increase (from four to eight to forty in fig. 4.1), the sampling distribution takes on the shape of a smooth curve with a peak at the true ratio of 3:1 (75% tall progeny)—that is, there is a high probability of getting very close to the true ratio. However, there is some chance the ratio will be fairly far off, and a very small part of the time our ratio will be very far off. It is important to see that any ratio could arise in a given experiment even though the true ratio is 3:1. At what point do we decide that an experimental result is not indicative of a 3:1 ratio?

Statisticians have agreed on a convention. When all the frequencies are plotted, as in figure 4.1, we can treat the area under the curve as one unit, and we can draw lines to mark 95% of this area (fig. 4.2). Any ratios in-

Table 4.1 Sampling Distribution for Sample Sizes of Four, Eight, and Forty, Given a 3:1 Ratio of Tall and Dwarf Plants

$n = 4$		$n = 8$		$n = 40$	
No. Tall Plants	Probability*	No. Tall Plants	Probability*	No. Tall Plants	Probability*
4	$\frac{81}{256} = 0.32$	8	0.10	40	0.00001
3	$\frac{108}{256} = 0.42$	7	0.27	39	0.0001
2	$\frac{54}{256} = 0.21$	6	0.31	38	0.0009
1	$\frac{12}{256} = 0.05$	5	0.21
0	$\frac{1}{256} = 0.004$	4	0.09	30	0.14
		3	0.02
		2	0.004	2	0.59×10^{-20}
		1	0.0004	1	0.10×10^{-21}
		0	0.00002	0	0.83×10^{-24}

*Probabilities are calculated from the binomial theorem.

$$\text{probability} = (n!/s!t!)p^s q^t$$

where n = number of progeny observed

s = number of progeny that are tall

t = number of progeny that are dwarf

p = probability of a progeny plant being tall (3/4)

q = probability of a progeny plant being dwarf (1/4)

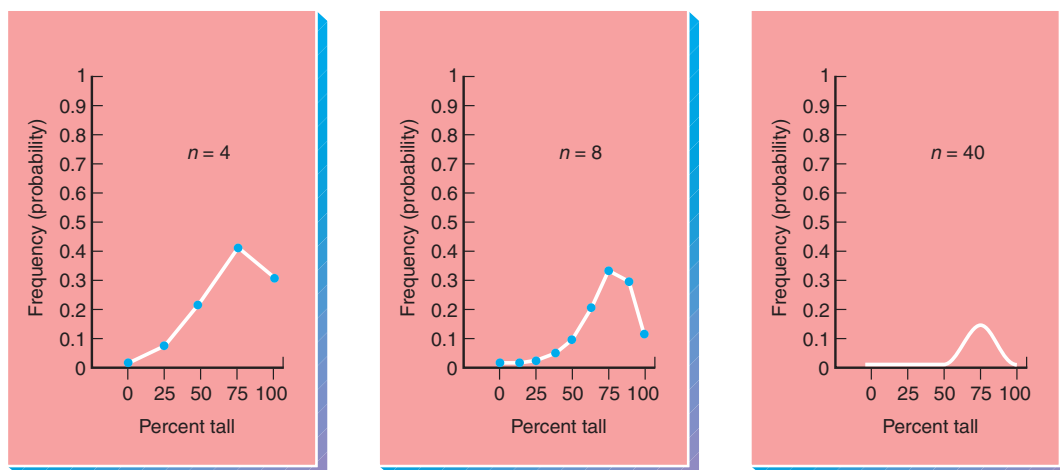


Figure 4.1 Sampling distributions from an experiment with an expected ratio of three tall to one dwarf plant. As the sample size, n , gets larger, the distribution curve becomes smoother. These distributions are plotted terms of the binomial expansion (see table 4.1). Note also that as n gets larger, the peak of the curve gets lower because as more points (possible ratios) are squeezed in along the x-axis, the probability of producing any one ratio decreases.

cluded within the 95% limits are considered supportive of (failing to reject) the hypothesis of a 3:1 ratio. Any ratio in the remaining 5% area is considered unacceptable. (Other conventions also exist, such as rejection within the outer 10% or 1% limits; we consider these at the end

of the chapter.) Thus, it is possible to see whether the experimental data support our hypothesis (in this case, the hypothesis of 3:1). One in twenty times (5%) we will make a **type I error**: We will reject a true hypothesis. (A **type II error** is failing to reject a false hypothesis.)

To determine whether to reject a hypothesis, we must derive a frequency distribution for each type of experiment. Mendel could have used the distribution shown in figure 4.1 for seed coat or seed color, as long as he was expecting a 3:1 ratio and had a similar sample size. What about independent assortment, which predicts a 9:3:3:1 ratio? A geneticist would have to calculate a new sampling distribution based on a 9:3:3:1 ratio and a particular sample size. Statisticians have devised short-cut methods by using standardized probability distributions. Many are in use, such as the *t*-distribution, binomial distribution, and chi-square distribution. Each is useful for particular kinds of data; geneticists usually use the chi-square distribution to test hypotheses regarding breeding data.

Chi-Square

When sample subjects are distributed among discrete categories such as tall and dwarf plants, geneticists frequently use the **chi-square distribution** to evaluate data. The formula for converting categorical experimental data to a chi-square value is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where χ is the Greek letter chi, O is the observed number for a category, E is the expected number for that category, and Σ means to sum the calculations for all categories.

A chi-square (χ^2) value of 0.60 is calculated in table 4.2 for Mendel's data on the basis of a 3:1 ratio. If Mendel had originally expected a 1:1 ratio, he would have calculated a chi-square of 244.45 (table 4.3). However, these χ^2 values have little meaning in themselves: they are not probabilities. We can convert them to probabilities by determining where the chi-square value falls in relation to the area under the chi-square distribution curve. We usually use a chi-square table that contains probabilities that have already been calculated (table 4.4). Before we can use this table, however, we must define the concept of **degrees of freedom**.

Reexamination of the chi-square formula and tables 4.2 and 4.3 reveals that each category of data contributes to the total chi-square value, because chi-square is a summed value. We therefore expect the chi-square value to increase as the total number of categories increases. That is, the more categories involved, the larger the chi-square value, even if the sample fits relatively well against the hypothesized ratio. Hence, we need some way of keeping track of categories. We can do this with degrees of freedom, which is basically a count of independent categories. With Mendel's data, the total number of offspring is 1,064, of which 787 had tall stems. Therefore, the

short-stem group had to consist of 277 plants (1,064 – 787) and isn't an independent category. For our purposes here, degrees of freedom equal the number of categories minus one. Thus, with two phenotypic categories, there is only one degree of freedom.

Table 4.4, the table of chi-square probabilities, is read as follows. Degrees of freedom appear in the left column. We are interested in the first row, where there is one degree of freedom. The numbers across the top of the table are the probabilities. We are interested in the next-to-the-last column, headed by the 0.05. We thus gain the following information from the table: *The probability is 0.05 of getting a chi-square value of 3.841 or larger by chance alone, given that the hypothesis is correct.* This statement formalizes the information in our discussion of frequency distributions. Hence, we are interested in how large a chi-square value will be found in the 5% unacceptable area of the curve. For Mendel's plant experiment, the **critical chi-square** (at $p = 0.05$, one degree of freedom) is 3.841. This is the value to which we compare the calculated χ^2 values (0.60 and 244.45). Since the chi-square value for the 3:1 ratio is 0.60 (table 4.2), which is less than the critical value of 3.841, we do not reject the hypothesis of a 3:1 ratio. But since χ^2 for the 1:1 ratio (table 4.3) is 244.45, which is greater than the critical value, we reject the hypothesis of a 1:1 ratio. Notice that once we did the chi-square test for the 3:1 ratio and failed to reject the hypothesis, no other statistical tests were needed: Mendel's data are consistent with a 3:1 ratio.

A word of warning when using the chi-square: If the expected number in any category is less than five, the conclusions are not reliable. In that case, you can repeat the experiment to obtain a larger sample size, or you can

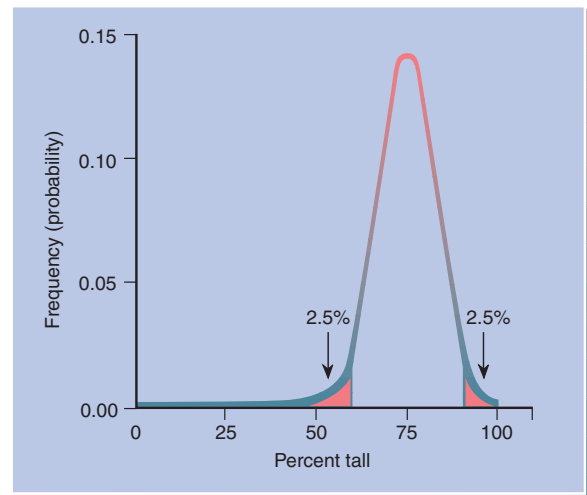


Figure 4.2 Sampling distribution of figure 4.1, $n = 40$. By convention, 5% of the area is marked off (2.5% at each end).

combine categories. Note also that chi-square tests are always done on whole numbers, not on ratios or percentages.

Failing to Reject Hypotheses

Hypothesis testing, in general, involves testing the assumption that there is no difference between the observed and the expected samples. Therefore, the hypothesis against which the data are tested is referred to as the **null hypothesis**. If the null hypothesis is not rejected, then we say that the data are consistent with it, not that the hypothesis has been proved. (As previously discussed, it is always possible we are not rejecting a false hypothesis or are rejecting the true one.) If, however, the

hypothesis is rejected, as we rejected a 1:1 ratio for Mendel’s data, we fail to reject the alternative hypothesis: that there is a difference between the observed and the expected values. We may then retest the data against some other hypothesis. (We don’t say “accept the hypothesis” but rather “fail to reject the hypothesis,” because supportive numbers could arise for many reasons. Our failure to reject is tentative acceptance of a hypothesis. However, we are on stronger ground when we reject a hypothesis.)

The use of the 0.05 probability level as a cutoff for rejecting a hypothesis is a convention called the **level of significance**. When a hypothesis is rejected at that level, statisticians say that the data depart *significantly* from the expected ratio. Other levels of significance are also

Table 4.2 Chi-Square Analysis of One of Mendel’s Experiments, Assuming a 3:1 Ratio

	Tall Plants	Dwarf Plants	Total
Observed numbers (<i>O</i>)	787	277	1,064
Expected ratio	3/4	1/4	
Expected numbers (<i>E</i>)	798	266	1,064
<i>O</i> – <i>E</i>	–11	11	
(<i>O</i> – <i>E</i>) ²	121	121	
(<i>O</i> – <i>E</i>) ² / <i>E</i>	0.15	0.45	0.60 = χ^2

Table 4.3 Chi-Square Analysis of One of Mendel’s Experiments, Assuming a 1:1 Ratio

	Tall Plants	Dwarf Plants	Total
Observed numbers (<i>O</i>)	787	277	1,064
Expected ratio	1/2	1/2	
Expected numbers (<i>E</i>)	532	532	1,064
<i>O</i> – <i>E</i>	255	–255	
(<i>O</i> – <i>E</i>) ²	65,025	65,025	
(<i>O</i> – <i>E</i>) ² / <i>E</i>	122.23	122.23	244.45 = χ^2

Table 4.4 Chi-Square Values

Degrees of Freedom	Probabilities						
	0.99	0.95	0.80	0.50	0.20	0.05	0.01
1	0.000	0.004	0.064	0.455	1.642	3.841	6.635
2	0.020	0.103	0.446	1.386	3.219	5.991	9.210
3	0.115	0.352	1.005	2.366	4.642	7.815	11.345
4	0.297	0.711	1.649	3.357	5.989	9.488	13.277
5	0.554	1.145	2.343	4.351	7.289	11.070	15.086
6	0.872	1.635	3.070	5.348	8.558	12.592	16.812
7	1.239	2.167	3.822	6.346	9.803	14.067	18.475
8	1.646	2.733	4.594	7.344	11.030	15.507	20.090
9	2.088	3.325	5.380	8.343	12.242	16.919	21.666
10	2.558	3.940	6.179	9.342	13.442	18.307	23.209
15	5.229	7.261	10.307	14.339	19.311	24.996	30.578
20	8.260	10.851	14.578	19.337	25.038	31.410	37.566
25	11.524	14.611	18.940	24.337	30.675	37.652	44.314
30	14.953	18.493	23.364	29.336	36.250	43.773	50.892

used, such as 0.01. If a calculated chi-square is greater than the critical value in the table at the 0.01 level, we say that the data depart in a *highly significant* manner from the null hypothesis. Since the chi-square value at the 0.01 level is larger than the value at the 0.05 level, it is more difficult to reject a hypothesis at this level and hence more convincing when it is rejected. Other levels of rejection are also set. In clinical trials of medication, for example, experimenters attempt to make it very easy to reject the null hypothesis: a level of significance of 0.10 or higher is set. The rationale is that it is not desirable to

discard a drug or treatment that may well be beneficial. Since the null hypothesis states that the drug has no effect—that is, the control and drug groups show the same response—clinicians would rather be overly conservative. Not rejecting the hypothesis means concluding that the drug has no effect. Rejecting the hypothesis means that the drug has some effect and should be tested further. It is much better to have to retest some drugs that are actually worthless than to discard drugs that have potential value.

S U M M A R Y

STUDY OBJECTIVE 1: To understand the rules of probability and how they apply to genetics 71–74

We have examined the rules of probability theory relevant to genetic experiments. Probability theory allows us to predict the outcomes of experiments. The probability (P) of independent events occurring is calculated by multiplying their separate probabilities. The probability of mutually exclusive events occurring is calculated by adding their individual probabilities. And the probability of unordered events is defined by the polynomial expansion $(p + q + r + \dots)^n$:

$$P = \frac{n!}{s!t!u! \dots} p^s q^t r^u \dots$$

STUDY OBJECTIVE 2: To understand the use of the chi-square statistical test in genetics 74–78

To assess whether data gathered during an experiment actually support a particular hypothesis, it is necessary to determine what the probability is of getting a particular data set when the null hypothesis is correct. One way to do this is through the chi-square test:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

This test gives us a method of quantifying the confidence we can place in the results obtained from typical genetic experiments. The rules of probability and statistics allow us to devise hypotheses about inheritance and to test these hypotheses with experimental data.

S O L V E D P R O B L E M S

PROBLEM 1: Mendel self-fertilized a dihybrid plant that had round, yellow peas. In the offspring generation: What is the probability that a pea picked at random will be round and yellow? What is the probability that five peas picked at random will be round and yellow? What is the probability that of five peas picked at random, four will be round and yellow, and one will be wrinkled and green?

Answer: The offspring peas will be round and yellow, round and green, wrinkled and yellow, and wrinkled and green in a ratio of 9:3:3:1. Thus, the probabilities that a pea picked at random will be one of these four categories are 9/16, 3/16, 3/16, and 1/16, respectively. Thus, the probability that a pea picked at random will be round

and yellow is 9/16, or 0.563. The probability of picking five of these peas in a row is $(9/16)^5$ or 0.056. The probability that of five peas picked at random, four will be round and yellow, and one will be wrinkled and green, is (substituting into the binomial equation): $(5!/4!1!)(9/16)^4 (1/16)^1 = 5(9^4)/(16^5) = 5(0.006) = 0.031$.

PROBLEM 2: On a chicken farm, walnut-combed fowl that were crossed with each other produced the following offspring: walnut-combed, 87; rose-combed, 31; pea-combed, 30; and single-combed, 12. What hypothesis might you have (based on chapter 2) about the control of comb shape in fowl? Do the data support that hypothesis?

Answer: The numbers 87, 31, 30, and 12 are very similar to 90, 30, 30, and 10, which would be a perfect fit to a 9:3:3:1 ratio. We might expect that ratio, having previously learned something about how comb type is inherited in fowl. Thus, we hypothesize that inheritance of comb type is by two loci, and that dominant alleles at both result in walnut combs, a dominant allele at one locus and recessives at the other result in rose or pea combs, and the recessive alleles at both loci result in a single comb. The results of the cross of dihybrids should produce fowl with the four comb types in a 9:3:3:1 ratio of walnut-, rose-, pea-, and single-combed fowl, respectively. Therefore, our observed numbers are 87, 31, 30, and 10 (sum = 160). Our expected ratio is 9:3:3:1, or 90, 30, 30, and 10 fowl, which are 9/16, 3/16, 3/16, and 1/16, respectively, of the sum of 160. We therefore set up the following chi-square table:

	Comb Type				Total
	Walnut	Rose	Pea	Single	
Observed Numbers (<i>O</i>)	87	31	30	12	160
Expected Ratio	9/16	3/16	3/16	1/16	
Expected Numbers (<i>E</i>)	90	30	30	10	160
<i>O</i> - <i>E</i>	-3	1	0	2	
(<i>O</i> - <i>E</i>) ²	9	1	0	4	
(<i>O</i> - <i>E</i>) ² / <i>E</i>	0.1	0.033	0	0.4	0.533 = χ^2

There are three degrees of freedom since there are four categories of combs. ($4 - 1 = 3$). The critical chi-square value with three degrees of freedom and probability of 0.05 = 7.815 (table 4.4). Since our calculated chi-square value (0.533) is less than this critical value, we cannot reject our hypothesis. In other words, our data are consistent with the hypothesis of a 9:3:3:1 phenotypic ratio, indicative of a two-locus genetic model with dominance at each locus.

E X E R C I S E S A N D P R O B L E M S *

PROBABILITY

- Assuming a 1:1 sex ratio, what is the probability that five children produced by the same parents will consist of
 - three daughters and two sons?
 - alternating sexes, starting with a son?
 - alternating sexes?
 - all daughters?
 - all the same sex?
 - at least four daughters?
 - a daughter as the eldest child and a son as the youngest?
 (See also USE OF RULES)
- Phenylthiocarbamide (PTC) tasting is dominant (*T*) to nontasting (*t*). If a taster woman with a nontaster father produces children with a taster man, and the man previously had a nontaster daughter, what would be the probability that
 - their first child would be a nontaster?
 - their first child would be a nontaster girl?
 - if they had six children, they would have two nontaster sons, two nontaster daughters, and two taster sons?
 - their fourth child would be a taster daughter?
 (See also USE OF RULES)

- Albinism is recessive; assume for this problem that blue eyes are also recessive (albinos have blue eyes). What is the probability that two brown-eyed persons, heterozygous for both traits, would produce (remembering epistasis)
 - five albino children?
 - five albino sons?
 - four blue-eyed daughters and a brown-eyed son?
 - two sons genotypically like their father and two daughters genotypically like their mother?
 (See also USE OF RULES)
- On the average, about one child in every ten thousand live births in the United States has phenylketonuria (PKU). What is the probability that
 - the next child born in a Boston hospital will have PKU?
 - after that child with PKU is born, the next child born will have PKU?
 - two children born in a row will have PKU?
- In fruit flies, the diploid chromosome number is eight.
 - What is the probability that a male gamete will contain only paternal centromeres or only maternal centromeres?

- b. What is the probability that a zygote will contain only centromeres from male grandparents? (Disregard the problems that the sex chromosomes may introduce.)
6. How many seeds should Mendel have tested to determine with complete certainty that a plant with a dominant phenotype was heterozygous? With 99% certainty? With 95% certainty? With "pretty reliable" certainty?
7. What chance do a man and a woman have of producing one son and one daughter?
8. PKU and albinism are two autosomal recessive disorders, unlinked in human beings. If two people, each heterozygous for both traits, produce a child, what is the chance of their having a child with
- PKU?
 - either PKU or albinism?
 - both traits?
9. In human beings, the absence of molars is inherited as a dominant trait. If two heterozygotes have four children, what is the probability that
- all will have no molars?
 - three will have no molars and one will have molars?
 - the first two will have molars and the second two will have no molars?
- (See also USE OF RULES)
10. Galactosemia is inherited as a recessive trait. If two normal heterozygotes produce children, what is the chance that
- one of four children will be affected?
 - three children will be born in this order: normal boy, affected girl, affected boy?
- (See also USE OF RULES)
11. A normal man (A) whose grandfather had galactosemia and a normal woman (B) whose mother was galactosemic want to produce a child. What is the probability that their first child will be galactosemic?
12. A city had nine hundred deaths during the year, and of these, three hundred were from cancer and two hundred from heart disease. What is the probability that the next death will be from
- cancer?
 - either cancer or heart disease?
13. A plant that has the genotype $AA\ bb\ cc\ DD\ EE$ is mated with one that is $aa\ BB\ CC\ dd\ ee$. F_1 individuals are selfed. What is the chance of getting an F_2 plant whose genotype exactly matches the genotype of one of the parents?

USE OF RULES

14. The ability to taste phenylthiocarbamide is dominant in human beings. If a heterozygous taster mates with a nontaster, what is the probability that of their five children, only one will be a taster?
15. In mice, coat color is determined by two independent genes, A and C, as indicated here: $A-C$, agouti; aaC , black; $A-c^a c^a$ and $aac^a c^a$, albino. If the following two mice are crossed $AaCc^a \times Aac^a c^a$, what is the probability that among the first six offspring, two will be agouti, two will be black, and two will be albino?

STATISTICS

16. The following data are from Mendel's original experiments. Suggest a hypothesis for each set and test this hypothesis with the chi-square test. Do you reach different conclusions with different levels of significance?
- Self-fertilization of round-seeded hybrids produced 5,474 round seeds and 1,850 wrinkled ones.
 - One particular plant from a yielded 45 round seeds and 12 wrinkled ones.
 - Of the 565 plants raised from F_2 round-seeded plants, 372, when self-fertilized, gave both round and wrinkled seeds in a 3:1 proportion, whereas 193 yielded only round seeds.
 - A violet-flowered, long-stemmed plant was crossed with a white-flowered, short-stemmed plant, producing the following offspring:
47 violet, long-stemmed plants
40 white, long-stemmed plants
38 violet, short-stemmed plants
41 white, short-stemmed plants
17. Mendel self-fertilized pea plants with round and yellow peas. In the next generation he recovered the following numbers of peas:
315 round and yellow peas
108 round and green peas
101 wrinkled and yellow peas
32 wrinkled and green peas
What is your hypothesis about the genetic control of the phenotype? Do the data support this hypothesis?
18. Two agouti mice are crossed, and over a period of a year they produce 48 offspring with the following phenotypes:
28 agouti mice
7 black mice
13 albino mice

What is your hypothesis about the genetic control of coat color in these mice? Do the data support that hypothesis?

19. Two curly-winged flies, when mated, produce sixty-one curly and thirty-five straight-winged progeny. Use a chi-square test to determine whether these numbers fit a 3:1 ratio.
20. A short-winged, dark-bodied fly is crossed with a long-winged, tan-bodied fly. All the F_1 progeny are

long-winged and tan-bodied. F_1 flies are crossed among themselves to yield 84 long-winged, tan-bodied flies; 27 long-winged dark-bodied flies; 35 short-winged, tan-bodied flies; and 14 short-winged, dark-bodied flies.

- a. What ratio do you expect in the progeny?
- b. Use the chi-square test to evaluate your hypothesis. Is the observed ratio within the expected range?

CRITICAL THINKING QUESTIONS

1. A friend shows you three closed boxes, one of which contains a prize, and asks you to choose one. Your friend then opens one of the two remaining boxes, a box she knows is empty. At that point, she gives you the opportunity to change your choice to the last remaining box. Should you?
2. If all couples wanted at least one child of each sex, approximately what would the average family size be?

Suggested Readings for chapter 4 are on page B-2.