

# 29

## HARDY-WEINBERG EQUILIBRIUM

### Objectives

- Understand the Hardy-Weinberg principle and its importance.
- Understand the chi-square test of statistical independence and its use.
- Determine the genotype and allele frequencies for a population of 1000 individuals.
- Use a chi-square test of independence to determine if the population is in Hardy-Weinberg equilibrium.
- Determine the genotypes and allele frequencies of an offspring population.

*Suggested Preliminary Exercises: Statistical Distributions; Hypothesis Testing*

### INTRODUCTION

When you picture all the breeds of dogs in the world—poodles, shepherds, retrievers, spaniels, and so on—it can be hard to believe they are all members of the same species. What accounts for their different appearance and talents, and how do dog breeders match up a male and female of a certain breed to produce prize-winning offspring? The physical and behavioral traits we observe in nature, such as height and weight, are known as the **phenotype**. An individual's phenotype is the product of its **genotype** (genetic make-up), or its environment, or both. In this exercise, we focus on the genetic make-up of a population and how it changes over time. This field of study is known as **population genetics**.

### *Genes, Alleles, and Genotypes*

A **gene**, loosely speaking, is a physical entity that is transmitted from parents to offspring and determines or influences traits (Hartl 2000). In one of the great achievements of the life sciences, Gregor Mendel studied the inheritance of flower color and seed shape in common peas and hypothesized the existence and behavior of such an entity of heredity many years before genes were actually described and shown to exist (Mendel 1866).

The multitude of genes in an organism reside on its chromosomes. A particular gene will be located at the same position, called the **locus** (plural, loci), on the

chromosomes of every individual in the populations. In sexually reproducing diploid organisms, individuals have two copies of each gene at a given locus; one copy is inherited paternally (from the father), the other maternally (from the mother). The two copies considered together determine the individual's genotype. Genes can exist in different forms, or states, and these alternative forms are called **alleles**. If the two alleles in an individual are identical, the individual's genotype is said to be **homozygous**. If the two are different, the genotype is **heterozygous**.

Although *individuals* are either homozygous or heterozygous at a particular locus, *populations* are described by their **genotype frequencies** and **allele frequencies**. The word *frequency* in this case means occurrence in a population. To obtain the genotype frequencies of a population, simply count up the number of each kind of genotype in the population and divide by the total number of individuals in the population. For example, if we study a population of 55 individuals, and 8 individuals are  $A_1A_1$ , 35 are  $A_1A_2$ , and 12 are  $A_2A_2$ , the genotype frequencies ( $f$ ) are

$$f(A_1A_1) = 8/55 = 0.146$$

$$f(A_1A_2) = 35/55 = 0.636$$

$$f(A_2A_2) = 12/55 = 0.218$$

$$\text{Total} = 1.00$$

The total of the genotype frequencies of a population always equals 1.

Allele frequencies, in contrast, describe the proportion of all alleles in the population that are of a specific type (Hartl 2000). For our population of 55 individuals above, there are a total of 110 alleles (of any kind) present in the population (each individual has two copies of a gene, so there are  $55 \times 2 = 110$  total alleles in the population). To calculate the allele frequencies of a population, we need to calculate how many alleles are  $A_1$  and how many are  $A_2$ . To calculate how many copies are  $A_1$ , we count the number of  $A_1A_1$  homozygotes and multiply that number by 2 (each homozygote has two  $A_1$  copies), then add to it the number of  $A_1A_2$  heterozygotes (each heterozygote has a single  $A_1$  copy). The total number of  $A_1$  copies in the population is then divided by the total number of alleles in the population to generate the **allele frequency**. The total number of  $A_1$  alleles in our example population is thus  $(2 \times 8) + (1 \times 35) = 51$ . The frequency of  $A_1$  is calculated as  $51/(2 \times 55) = 51/110 = 0.464$ . Similarly, the total number of  $A_2$  alleles in the population is  $(2 \times 12) + (1 \times 35) = 59$ , and the frequency of  $A_2$  is  $59/(2 \times 55) = 59/110 = 0.536$ .

As with genotype frequencies, the total of the allele frequencies of a population always equals 1. By convention, frequencies are designated by letters. If there are only two alleles in the population, these letters are conventionally  $p$  and  $q$ , where  $p$  is the frequency of one kind of allele and  $q$  is the frequency of the second kind of allele. For genes that have only two alleles,

$$p + q = 1 \qquad \text{Equation 1}$$

If there were more than two kinds of alleles for a particular gene, we would calculate allele frequencies for the other kinds of alleles in the same way. For example, if three alleles were present,  $A_1$ ,  $A_2$ , and  $A_3$ , the frequencies would be  $p$  (the frequency of the  $A_1$  allele),  $q$  (the frequency of the  $A_2$  allele) and  $r$  (the frequency of the  $A_3$  allele). No matter how many alleles are present in the population, the frequencies should always add to 1. In this exercise, we will keep things simple and focus on a gene that has only two alleles.

In summary, for a population of  $N$  individuals, the number of  $A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  genotypes are  $N_{A_1A_1}$ ,  $N_{A_1A_2}$ , and  $N_{A_2A_2}$ , respectively. If  $p$  represents the frequency of the  $A_1$  allele, and  $q$  represents the frequency of the  $A_2$  allele, the estimates of the allele frequencies in the population are

$$f(A_1) = p = (2N_{A_1A_1} + N_{A_1A_2})/2N \quad \text{Equation 2}$$

$$f(A_2) = q = (2N_{A_2A_2} + N_{A_1A_2})/2N \quad \text{Equation 3}$$

### The Hardy-Weinberg Principle

Population geneticists are not only interested in the genetic make-up of populations, but also how genotype and allele frequencies *change* from generation to generation. In the broadest sense, **evolution** is defined as the change in allele frequencies in a population over time (Hartl 2000). The Hardy-Weinberg principle, developed by G. H. Hardy and W. Weinberg in 1908, is the foundation for the genetic theory of evolution (Hardy 1908). It is one of the most important concepts that you will learn about in your studies of population biology and evolution.

Broadly stated, the Hardy-Weinberg principle says that given the initial genotype frequencies  $p$  and  $q$  for two alleles in a population, after a single generation of random mating the genotype frequencies of the offspring will be  $p^2:2pq:q^2$ , where  $p^2$  is the frequency of the  $A_1A_1$  genotype,  $2pq$  is the frequency of the  $A_1A_2$  genotype, and  $q^2$  is the frequency of the  $A_2A_2$  genotype. The sum of the genotype frequencies, as always, will sum to one; thus,

$$p^2 + 2pq + q^2 = 1 \quad \text{Equation 4}$$

This equation is the basis of the Hardy-Weinberg principle.

The Hardy-Weinberg principle further predicts that genotype frequencies and allele frequencies will remain constant in any succeeding generations—in other words, the frequencies will be in **equilibrium** (unchanging). For example, in a population with an  $A_1$  allele frequency  $p$  of 0.75 and an  $A_2$  allele frequency  $q$  of 0.25, in Hardy-Weinberg equilibrium, the genotype frequencies of the population should be:

$$f(A_1A_1) = p^2 = p \times p = 0.75 \times 0.75 = 0.5625$$

$$f(A_1A_2) = 2 \times p \times q = 2 \times 0.75 \times 0.25 = 0.375$$

$$f(A_2A_2) = q^2 = q \times q = 0.25 \times 0.25 = 0.0625$$

Now let's suppose that this founding population mates at random. The Hardy-Weinberg principle tells us that after just one generation of random mating, the genotype frequencies in the next generation will be

$$f(A_1A_1) = p^2 = p \times p = 0.75 \times 0.75 = 0.5625$$

$$f(A_1A_2) = 2 \times p \times q = 2 \times 0.75 \times 0.25 = 0.375$$

$$f(A_2A_2) = q^2 = q \times q = 0.25 \times 0.25 = 0.0625$$

Additionally, the initial allele frequencies will remain at 0.75 and 0.25. These frequencies (allele and genotype) will remain unchanged over time.

The Hardy-Weinberg principle is often called the “null model of evolution” because genotypes and allele frequencies of a population in Hardy-Weinberg equilibrium will remain unchanged over time. That is, populations won't evolve. When populations violate the Hardy-Weinberg predictions, it suggests that some evolutionary force is acting to keep the population out of equilibrium. Let's walk through an example.

Suppose a population is founded by 3,000  $A_1A_1$  and 1,000  $A_2A_2$  individuals. From Equation 2, the frequency of the  $A_1$  allele,  $p$ , is  $(2 \times 3000 + 0)/(2 \times 4000) = 0.75$ . Because  $p + q$  must equal 1,  $q$  must equal  $1 - p$ , or 0.25. So, since  $p$  and  $q$  are equal to the values we used above to calculate the equilibrium genotype frequencies, if this population were in Hardy-Weinberg equilibrium, 56% of the population *should* be homozygous  $A_1A_1$ , 38% *should* be heterozygous, and 6% *should* be homozygous  $A_2A_2$ . But the *actual* genotype frequencies in this population are 75% homozygous  $A_1A_1$  and 25% homozygous

$A_2A_2$ —there are *no* heterozygotes! So this founding population is *not* in Hardy-Weinberg equilibrium.

To determine whether an observed population’s deviations from Hardy-Weinberg expectations might be due to random chance, or whether the deviations are so significant that we must conclude, as we did in the preceding example, that the population is not in equilibrium, we perform a statistical test.

**The Chi-Square Test of Independence**

Once you know the actual allele frequencies *observed* in your population and the genotype frequencies you *expected* to see in an equilibrium population, you have the information to answer the question, “Is the population in fact in a state of Hardy-Weinberg equilibrium?”

When we know the values of what we expected to observe and what we actually observed, a **chi-square ( $\chi^2$ ) test of independence** is commonly used to determine whether the observed values in fact match the expected value (the **null model** or **null hypothesis**) or whether the observed values deviate significantly from what we expect to find (in which case we reject the null model).

Chi-square statistical tests are performed to test hypotheses in all the life and social sciences. The test basically asks whether the differences between observed and expected values could be due to chance. The mathematical basis of the test is the equation

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad \text{Equation 5}$$

where  $O$  is the observed value,  $E$  is the expected value, and  $\Sigma$  means you sum the values for different observations. Hardy-Weinberg genotype frequencies offer a good opportunity to use the chi-square test.

In conducting a  $\chi^2$  test of independence, it’s useful to set up your data in a table format, where the observed values go in the top row of the table, and the expected values go in row 2. The expected values for each genotype are those predicted by Hardy-Weinberg, computed as  $p^2 \times N$ ,  $2pq \times N$ , and  $q^2 \times N$  for the  $A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  genotypes, respectively. If  $N = 1000$  individuals and  $p = 0.5$  and  $q = 0.5$ , our expected numbers would be 250  $A_1A_1$ , 500  $A_1A_2$ , and 250  $A_2A_2$  (Figure 1).

To compute the  $\chi^2$  test statistic, we start by computing the difference between the observed and expected numbers for a genotype, square this difference, and then divide by the expected number for that genotype. We do this for the remaining genotypes, and then add the terms together:

$$\chi^2 = \frac{(O_{A1A1} - E_{A1A1})^2}{E_{A1A1}} + \frac{(O_{A1A2} - E_{A1A2})^2}{E_{A1A2}} + \frac{(O_{A2A2} - E_{A2A2})^2}{E_{A2A2}}$$

	J	K	L	M
7	<b>Parental Population</b>			
8	<b>A2A1</b>			
9	<b>A1A1</b>	<b>A1A2</b>	<b>A2A2</b>	
10	<b>Observed</b>	258	504	238
11	<b>Expected</b>	$p^2 * N = 250$	$2pq * N = 500$	$q^2 * N = 250$

**Figure 1** The top row gives the observed genotypes in a population of 1,000 individuals in which both  $p$  and  $q = 0.5$ . The bottom row gives the expected genotype distribution for those values of  $p$  and  $q$  if the population were in Hardy-Weinberg equilibrium.

The  $\chi^2$  test statistic for Figure 1 would be computed as

$$\chi^2 = \frac{(258 - 250)^2}{250} + \frac{(504 - 500)^2}{500} + \frac{(238 - 250)^2}{250} = 0.864$$

### D.F. and Critical Value

You now need to see where your computed  $\chi^2$  test statistic falls on the **theoretical  $\chi^2$  distribution**. If you are familiar with the normal distribution, you know that the mean and standard deviation control the shape and placement of the distribution on the  $x$ -axis (see Exercise 3, "Statistical Distributions"). A  $\chi^2$  distribution, in contrast, is characterized by a parameter called **degrees of freedom (d.f.)**, which controls the shape of the theoretical  $\chi^2$  distribution. The degrees of freedom value is computed as

$$\text{d.f.} = (\text{number of rows minus } 1) \times (\text{number of columns minus } 1)$$

or

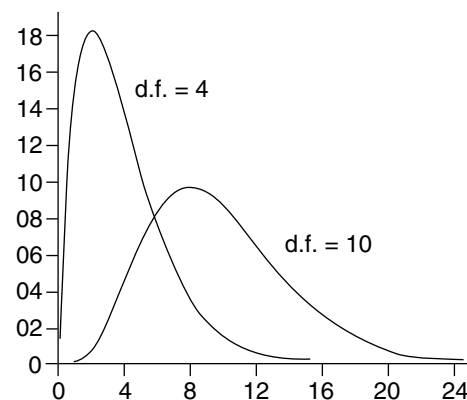
$$\text{d.f.} = (r - 1) \times (c - 1) \quad \text{Equation 6}$$

In Figure 1, we had two rows (observed and expected) and three columns (three kinds of genotypes), so our degrees of freedom =  $(2 - 1) \times (3 - 1) = 2$ .

The mean of a  $\chi^2$  distribution is its degrees of freedom, and the mode of a  $\chi^2$  distribution is the degrees of freedom minus 2. The distribution has a positive skew, but this skew diminishes as the degrees of freedom increases. Figure 2 shows two  $\chi^2$  distributions for different degrees of freedom. The  $\chi^2$  distributions in Figure 2 were generated from an infinite number of  $\chi^2$  tests performed on data sets where *no effects* were present. In other words, the theoretical  $\chi^2$  distribution is a null distribution. Even when no effects are present, however, you can see that, by chance, some  $\chi^2$  test statistics are large and appear with a low frequency. Thus, you can get a very large test statistic by chance even when there is no effect.

By convention, we are interested in knowing if our computed  $\chi^2$  statistic is larger than 95% of the statistics from the theoretical curve. The 95% value of the theoretical curve's  $\chi^2$  statistic is called the **critical  $\chi^2$  value**, and at this value, exactly 5% of the test statistics in the  $\chi^2$  distribution are greater than this critical value ( $\alpha = 0.05$ ; see Exercise 5, "Hypothesis Testing"). For example, the critical value for a  $\chi^2$  distribution with 4 degrees of freedom is 9.49, which means that 5% of the test statistics in the  $\chi^2$  distribution are equal to or greater than this value. The critical value for a  $\chi^2$  distribution with 10 degrees of freedom is 18.31.

Table 1 gives the critical values for  $\chi^2$  distributions with various degrees of freedom when  $\alpha = 0.05$  (the "95% confidence level"). Tables of  $\chi^2$  critical values for different  $\alpha$  values can be found in almost any statistics text. If our computed statistic is *less* than the



**Figure 2** Two  $\chi^2$  distributions. Note that the curve steepens (positive skew increases) when the degrees of freedom (d.f.) parameter is smaller.

critical value, we conclude that any difference between our observed and expected values are not **significant**—the difference could be due to chance—and we accept the null hypothesis (i.e., that the population is in Hardy-Weinberg equilibrium). But if our computed statistic is *greater* than the critical value, we conclude that the difference *is* significant, and we reject the null model (i.e., we conclude the population is not in equilibrium).

TABLE 1. Critical values of  $\chi^2$  at the 0.05 level of significance ( $\alpha$ )

Degrees of freedom	$\alpha = 0.05$	Degrees of freedom	$\alpha = 0.05$
1	3.84	11	19.68
2	5.99	12	21.03
3	7.82	13	22.36
4	9.49	14	23.69
5	11.07	15	25.00
6	12.59	16	26.30
7	14.07	17	27.59
8	15.51	18	28.87
9	16.92	19	30.14
10	18.31	20	31.41

Source:  $\chi^2$  values from R. A. Fisher and F. Yates, 1938, *Statistical Tables for Biological, Agricultural, and Medical Research*. Longman Group Ltd., London.

How do you interpret a significant  $\chi^2$  test? Interpretation requires that you examine the observed and expected values and determine which genotypes affected the value of the computed  $\chi^2$  statistic the most. In general, the larger the deviation between the observed and expected values, the greater the genotype contributed to the  $\chi^2$  statistic. In our first example, in which we expected 38% of an equilibrium population would be heterozygotes but in fact observed *no* heterozygotes, the deviation from Hardy-Weinberg expectations is caused primarily by the absence of heterozygotes. You could then proceed to form hypotheses as to *why* there are no heterozygotes.

What forces might keep a population out of Hardy-Weinberg equilibrium? **Evolutionary forces** include natural selection, genetic drift, gene flow, nonrandom mating (inbreeding), and mutation. These forces are introduced in other exercises, but here we will set up the “null model” of a population in Hardy-Weinberg equilibrium.

## PROCEDURES

In this exercise, you will develop a spreadsheet model of a single gene with two alleles in population and will explore various properties of Hardy-Weinberg equilibrium.

## INSTRUCTIONS

A. Set up the model parent population.

## ANNOTATION

Here we are concerned with a single locus, and imagine that this locus has two alleles,  $A_1$  and  $A_2$ . Thus, an individual can be homozygous  $A_1A_1$ , heterozygous  $A_1A_2$ , or homozygous  $A_2A_2$  at the locus.

1. Open a new spreadsheet and set up titles and column headings as shown in Figure 3.

	A	B	C	D	E	F	G	H
1	<b>Hardy-Weinberg Equilibrium</b>							
2								
3	Allele	$p = A1 =$			Calculated	$p =$		
4	frequencies	$q = A2 =$			frequencies	$q =$		
5								
6		Parental		Random	Mom's	Random	Dad's	Offspring
7	Individual	genotype	Gamete	mom	egg	dad	sperm	genotype

Figure 3

2. Set up a linear series from 0 to 999 to represent 1000 individuals in cells A8–A1007.

3. In cell C3, enter a value for  $p$ .

4. In cell C4, enter a formula to compute the value for  $q$ .

5. In cells B8–B1007, enter an **IF** formula to assign genotypes to each individual in the population based on the allele frequencies designated in cells C3 and C4.

6. Set up new spreadsheet headings as shown in Figure 4.

In cell A8, enter the value 0.

In cell A9, enter  $=A8+1$ . Copy the formula in cell A9 down to cell 1007 to designate the 1,000 individuals in the population.

Enter 0.5 in cell C3 to indicate that the frequency of the  $A_1$  allele, or  $p$ , is 0.5.

Enter the formula  $=1-C3$  in cell C4 to designate the frequency of the  $A_2$  allele, or  $q$ . Remember that  $p + q = 1$ .

Enter the formula  $=IF(RAND()<C3,"A1","A2")\& IF(RAND()<C3,"A1","A2")$  in cell B8. Copy this formula down to cell B1007.

The **IF** formula returns one value if a condition you specify is true, and another value if the condition you specify is false. The **RAND()** part of the formula in cell B8 tells the spreadsheet to choose a random number between 0 and 1. Then, *if* that random number is less than the value designated in cell C3, assign it an allele of  $A_1$ ; otherwise, assign it a value of  $A_2$ . Because there are two alleles for a given locus, you need to repeat the formula again, and then join the alleles obtained from the two **IF** formulas by using the **&** symbol. Once you've obtained genotypes for individual 1, copy this formula down to cell B1007 to obtain genotypes for all 1,000 individuals in the population.

	J	K	L	M
7	<b>Parental Population</b>			
8			<b>A1A2</b>	
9		<b>A1A1</b>	<b>A2A1</b>	<b>A2A2</b>
10	<b>Observed</b>			
11	<b>Expected</b>			
12				
13	Hand-calculated chi-square			
14	Degrees of freedom			
15	Chi test statistic			
16	Spreadsheet-calculated chi-square			
17	Significantly different from H-W prediction?			

Figure 4

7. In cells K10, L10, and M10, use the **COUNTIF** formula to count the number of  $A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  genotypes.

8. In cell G3, enter a formula to calculate the actual frequency of the  $A_1$  allele. In cell G4, enter a formula to calculate the actual frequency of the  $A_2$  allele.

9. Save your work.

**B. Calculate expected genotype frequencies in the parent population.**

1. In cell K11, enter a formula to calculate the expected number of  $A_1A_1$  genotypes, given the  $p$  value calculated in cell G3.

2. Calculate the expected number of heterozygotes in cell L11.

3. Calculate the expected number of  $A_2A_2$  genotypes in cell M11.

4. Graph your observed and expected results.

The **COUNTIF** formula counts the number of cells within a range that meet the given criteria. It has the syntax **COUNTIF(range,criteria)**, where range is the range of cells from which you want to count cells, and criteria is what you want to count. We used the formulae:

- Cell K10 =**COUNTIF(\$B\$8:\$B\$1007,"A1A1")**
- Cell L10  
=**COUNTIF(\$B\$8:\$B\$1007,"A1A2")+COUNTIF(\$B\$8:\$B\$1007,"A2A1")**
- Cell M10 =**COUNTIF(\$B\$8:\$B\$1007,"A2A2")**

The formula in cell K10 counts the number of  $A_1A_1$  individuals in cells B8 through B1007. In cell L10, you'll want to count both the  $A_1A_2$  and the  $A_2A_1$  heterozygotes. Your total observations should add to 1000. You can double-check this by entering =**SUM(K10:M10)** in cell N10.

The values from these formulae are your "observed" genotypes, and you'll compare these to the genotypes predicted by Hardy-Weinberg. (Your observed genotypes should be in Hardy-Weinberg equilibrium because of the way you assigned the genotypes. In a natural setting, however, you probably won't know the initial frequencies, but you can count genotypes, and then determine if the organisms are in Hardy-Weinberg equilibrium or not.)

Enter the formula =**(K10\*2+L10)/(2\*A1007)** in cell G3.

Enter the formula =**1-G3** in cell G4.

Since each individual carries two copies of each gene, your population of 1,000 individuals has 2,000 "gene copies" (alleles) present. To calculate the allele frequency, you simply calculate what proportion of those 2000 alleles are  $A_1$ , and what proportion are  $A_2$ . The frequency of the  $A_1$  allele is 2 times the number of  $A_1A_1$  genotypes, plus the  $A_1$ 's from the heterozygotes. The frequency of the  $A_2$  allele is 2 times the number of  $A_2A_2$  genotypes, plus the  $A_2$ 's from the heterozygotes. Since  $p + q = 1$ ,  $q$  can be computed also as  $1 - p$ . Your estimates of allele frequencies should add to 1.

Now that you have computed the observed allele frequencies, you can calculate the estimated genotype frequencies predicted by Hardy-Weinberg. Remember that if the population is in Hardy-Weinberg equilibrium, the genotype frequencies should be  $p^2 + 2pq + q^2$ . This means that the number of  $A_1A_1$  genotypes should be  $p \times p$  ( $p^2$ ), the number of  $A_1A_2$  genotypes should be  $2 \times p \times q$ , and the number of  $A_2A_2$  genotypes should be  $q \times q$  ( $q^2$ ).

Enter the formula =**\$G\$3^2\*1000** in cell K11.

The caret symbol (^) followed by the number 2 indicates that the value should be squared. Thus, we obtained expected number of  $A_1A_1$  genotypes by multiplying  $p \times p$ , which gives us a proportion, and then multiplied this proportion by 1,000 to give us the number of individuals out of 1,000 that are expected to be  $A_1A_1$  if the population is in Hardy-Weinberg equilibrium.

Enter the formula =**2\*\$G\$3\*\$G\$4\*1000** in cell L11.

Enter the formula =**\$G\$4^2\*1000** in cell M11.

The expected numbers should add to 1000. You can double-check this by entering =**SUM(K11:M11)** in cell N11.

Use a column graph and label your axes fully. Your graph may look a bit different than Figure 5, and that's fine.

5. Interpret your graph. Does your population appear to be in Hardy-Weinberg equilibrium?

6. Press F9, the calculate key, to generate new random numbers and hence new genotypes. Does your population still appear to be in equilibrium?

### C. Calculate chi-square test statistics and probability.

1. In cell M13, enter the formula to calculate your  $\chi^2$  test statistic. Refer to Equation 5.

2. In cell M14, enter a value for degrees of freedom.

3. In cell M15, use the **CHIDIST** function to determine the probability of obtaining your  $\chi^2$  statistic.

4. In cell M16, double-check your work by using the **CHITEST** function to calculate your test statistic, degrees of freedom, and probability.

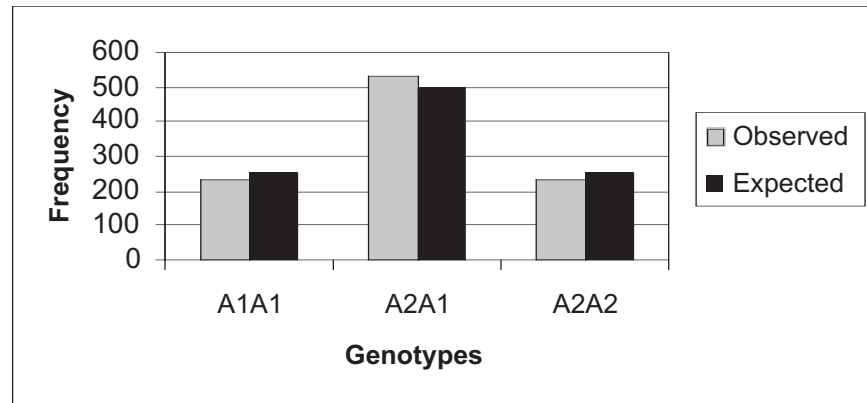


Figure 5

Now you are ready to perform a  $\chi^2$  test to verify whether your population's observed genotype frequencies are statistically similar to those predicted by Hardy-Weinberg.

Enter the formula `= (K10-K11)^2/K11+(L10-L11)^2/L11+(M10-M11)^2/M11` in cell M13. This corresponds to Equation 5:

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

Starting with  $A_1A_1$ , we observed 245 individuals and determined that there should be 255 individuals (you may have obtained slightly different numbers than that). Following the chi-square formula,  $245 - 255 = 10$ ,  $10^2 = 100$ , 100 divided by 255 = 0.392. Repeat this step for the  $A_1A_2$  and  $A_2A_2$  genotypes. As a final step, add your three calculated values together. This sum is your chi-square ( $\chi^2$ ) test statistic.

Enter the value 2 in cell M14.

Recall from Equation 6 that the degrees of freedom value is the (number of rows minus 1)  $\times$  (number of columns minus 1), or  $(r - 1) \times (c - 1)$ . In our example, we had two rows (observed and expected) and three columns (three kinds of genotypes), so our degrees of freedom =  $(2 - 1) \times (3 - 1) = 2$ .

Enter the formula `=CHIDIST(M13,M14)` in cell M15.

The **CHIDIST** function has the syntax **CHIDIST(x,degrees\_freedom)**, where **x** is the test statistic you want to evaluate and **degrees\_freedom** is the degrees of freedom for the test. The formula in cell M15 returns the probability of obtaining the test statistic you calculated, given the degrees of freedom—if this probability is less than 0.05, your test statistic exceeds the critical value. If this probability is greater than 0.05, your test statistic is less than the critical value. You can now make an informed decision as to whether your population is in Hardy-Weinberg equilibrium or not.

Enter the formula `=CHITEST(K10:M10,K11:M11)` in cell M16.

The **CHITEST** formula returns the test for independence (the probability) when you indicate the observed and expected values from a table. It has the syntax **CHITEST(actual\_range,expected\_range)**, where **actual\_range** is the range of observed data (in your case, cells K10–M10), and **expected\_range** is the range of expected data (in your case, cells K11–M11). This number should be very close to what you obtained in cell M15. (If it's not, you did something wrong.)

5. In cell M17, enter an **IF** formula to determine whether the probabilities you obtained in cell M15 is significant (i.e., significantly different from what would be expected by chance alone).

6. Answer questions 1 and 2 at the end of exercise before proceeding.

**D. Simulate random mating to produce the genotypes of the next ( $F_1$ ) generation.**

1. In cells C8–C1007, enter a formula to simulate the random assortment of alleles into gametes.

2. In cells D8 and F8, enter a formula to randomly select a male and a female from the population that will mate and produce a zygote.

Enter the formula **=IF(M15<0.05,"Yes","No")** in cell M17.

This **IF** formula tells the spreadsheet to evaluate the probability obtained in cell M15. By convention, if the value in M15 is *more* than 0.05, you would conclude that your observed frequencies are not significantly different than those expected by chance alone. If the value is *less* than 0.05, you would conclude that the population's observed genotypes are not in Hardy-Weinberg equilibrium.

Our results looked something like Figure 6 (your results are probably slightly different, and that's fine).

	J	K	L	M
7	<b>Parental Population</b>			
8			<b>A1A2</b>	
9		<b>A1A1</b>	<b>A2A1</b>	<b>A2A2</b>
10	<b>Observed</b>	253	498	249
11	<b>Expected</b>	252.004	499.992	248.004
12				
13	Hand-calculated chi-square			0.015872764
14	Degrees of freedom			2
15	Chi test statistic			0.992095028
16	Spreadsheet-calculated chi-square			0.992095028
17	Significantly different from H-W prediction?			No

**Figure 6**

Now that you have an idea of whether your population of 1,000 is in Hardy-Weinberg equilibrium, we will let your population mate and produce offspring that make up the next generation.

Enter the formula **=IF(RAND()<0.5,RIGHT(B8,2), LEFT(B8,2))** in cell C8. Copy this formula down to cell C1007.

Homozygotes can produce only one kind of gamete, while heterozygotes can produce both  $A_1$  and  $A_2$  gametes. We'll assume that each individual produces a single gamete, and that which of the two possible gametes are actually incorporated into the zygote is randomly determined. The formula in cell C8 tells the spreadsheet to draw a random number between 0 and 1 (the **RAND()** portion of the formula). If the random number is less than 0.5, the program returns the **RIGHT** two characters in cell B8; otherwise, it will return the **LEFT** two characters in cell B8. (You might want to explore the **RIGHT** and **LEFT** functions in more detail.) This formula simulates the random assortment of alleles into gametes that will ultimately fuse with another gamete to form a zygote.

Enter the formula **=ROUND(RAND()\*1000,0)** in cells D8 and F8. Copy the formula down to cells D1007 and F1007, respectively.

This formula simulates random mating by choosing a random female and random male from our population to mate. The formula tells the spreadsheet to draw a random number between 0 and 1, multiply this number by 1,000, then round it to 0 decimal places. This action will "choose" which individuals will mate. Note that not all individuals in

3. In columns E and G, enter **VLOOKUP** formulae to determine the gamete contributed by each parent randomly selected in step 2.

4. In cell H8, enter a formula to obtain the genotypes of the zygotes by pairing the egg and sperm alleles contributed by each parent.

**E. Calculate Hardy-Weinberg statistics for the  $F_1$  generation.**

1. Set up new column headings as shown in Figure 7.

the population will actually mate, but that each individual has the same probability of mating as every other individual in the population.

In cell E8 enter the formula =**VLOOKUP**(D8,\$A\$8:\$C\$1007,3). Copy this formula down to E1007.

In cell G8, enter the formula =**VLOOKUP**(F8,\$A\$8:\$C\$1007,3). Copy this formula down to G1007.

The formula in cell E8 tells the spreadsheet to look up the value in D8, which is the random mom, from the table A8 through A1007, and return the associated value listed in the third column of the table. In other words, find mom from column A and relay the gamete associated with that mom in column C. The formula in G8 does the same for the random dad.

The **VLOOKUP** function searches for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify in the table. It has the syntax **VLOOKUP**(lookup\_value,table\_array,col\_index\_num,range\_lookup), where **lookup\_value** is the value to be found in the first column of the table, **table\_array** is the table of information in which the data are looked up, and **col\_index\_num** is the column in the table that contains the value you want to return. **Range\_lookup** is either true or false. If **Range\_lookup** is not specified, by default it is set to "false," which indicates that an exact match will be found.

Enter the formula =**E8&G8** in cell H8. Copy this formula down to cell H1007.

Now you can determine if the offspring generation has genotypes predicted by Hardy-Weinberg. Remember, the Hardy-Weinberg principle holds that whatever the initial genotype frequencies for two alleles may be, after one generation of random mating, the genotype frequencies will be  $p^2:2pq:q^2$ . Additionally, both the genotype frequencies and the allele frequencies will remain constant in succeeding generations. The observed genotypes are calculated by tallying the different genotypes in cells H8–H1007. The expected genotypes are calculated based on the parental allele frequencies given in cells G3 and G4.

	J	K	L	M
20	<b>Offspring Population</b>			
21			<b>A1A2</b>	
22		<b>A1A1</b>	<b>A2A1</b>	<b>A2A2</b>
23	Observed			
24	Expected			
25				
26	Hand-calculated chi-square test statistic:			
27	Degrees of freedom			
28	Chi test statistic			
29	Spreadsheet-calculated chi-square			
30	Significantly different from H-W prediction?			

**Figure 7**

2. Enter formulae in cells K23–M24 to calculate observed and expected genotypes of the new generation.

3. Enter formulae in cells M26–M30 to determine if the new generation is in Hardy-Weinberg equilibrium.

4. Graph your observed and expected results.

If you've forgotten how to calculate a formula, refer to the formulas you entered for the parents as an aid. Double-check your results:

- K23 =COUNTIF(\$H\$8:\$H\$1007,"A1A1")
- L23 =COUNTIF(\$H\$8:\$H\$1007,"A1A2")+COUNTIF(\$H\$8:\$H\$1007,"A2A1")
- M23 =COUNTIF(\$H\$8:\$H\$1007,"A2A2")
- K24 =G\$3^2\*1000
- L24 =2\*G\$3\*G\$4\*1000
- M24 =G\$4^2\*1000

You can also simply copy and paste the formulae from the parental population; the program should automatically update your formulae to the new cells (but double-check, just to be sure).

## QUESTIONS

1. The Hardy-Weinberg model is often used as the "null model" for evolution. That is, when populations are out of Hardy-Weinberg equilibrium, it suggests that some kind of evolutionary process may be acting on the population. What are the assumptions of Hardy-Weinberg?
2. Press F9, the Calculate key, to generate a new set of random numbers, which in turn will generate new genotypes, new allele frequencies and new Hardy-Weinberg test statistics. Press F9 a number of times and track whether the population remains in Hardy-Weinberg equilibrium. Why, on occasion, will the population be out of HW equilibrium?
3. A basic tenet of the Hardy-Weinberg principle is that genotype frequencies of a population can be predicted if you know the allele frequencies. This allows you to answer such questions as *Under what allelic conditions should heterozygotes dominate the population?* In cell C3, modify the frequency of the  $A_1$  allele (the  $A_2$  allele will automatically be calculated). Begin with a frequency of 0, then increase its frequency by 0.1 until the frequency is 1. For *each* incremental value entered, record the expected genotype frequencies of  $A_1A_1$ ,  $A_1A_2$ , and  $A_2A_2$  given in cells K11–M11. (You can simply copy and paste these values into a new section of your spreadsheet, but make sure you use the Paste Values option to paste the expected genotypes.). Your spreadsheet might look something like this (but the frequencies will extend a few more rows until the frequency of  $A_1$  is 1:

	O	P	Q	R
13		<b>Expected genotypes</b>		
14	<b>Frequency of A1</b>	<b>A1A1</b>	<b>A1A2</b>	<b>A2A2</b>
15	0	0	0	1000
16	0.1	9	180	817
17	0.2	36	320	658
18	0.3	86	420	498
19	0.4	173	480	341
20	0.5	262	500	239

Make a graph of the relationship between frequency of the  $A_1$  allele (on the  $x$ -axis) and the expected numbers of genotypes. Use a line graph, and fully label your axes and give the graph a title. Consider the shapes of each curve, and write a one- or two-sentence description of the major points of the graph.

4. The Hardy-Weinberg principle states that after one generation of random mating, the genotype frequencies should be  $p^2:2pq:q^2$ . That is, even if a parental population is out of Hardy-Weinberg equilibrium, it should return to the equilibrium status after just one generation of random mating. Prove this to yourself by modifying the genotypes of the 1,000 individuals listed in column B. Let individuals 0–499 have genotypes  $A_1A_1$ ; individuals 500–999 have genotypes of  $A_2A_2$ . (You'll have to overwrite the formulas in those cells.) Estimate the gene frequencies and determine if this parental population is in Hardy-Weinberg equilibrium. Graph your results, and indicate the chi-square test statistic somewhere on your graph. After one generation of random mating, what are the allele frequencies and genotype frequencies? Is this "new" population in Hardy-Weinberg equilibrium?

### LITERATURE CITED

- Hardy, G. 1908. Mendelian proportions in a mixed population. *Science* 28: 49–50.
- Hartl, D. L. 2000. *A Primer of Population Genetics*, 3rd Edition. Sinauer Associates, Sunderland, MA.
- Mendel, G. 1866. Experiments in plant hybridization. Translated and reprinted in J. A. Peters (ed.), 1959. *Classic Papers in Genetics*. Prentice-Hall, Englewood Cliffs, NJ.