# 43

# *EFFECTIVE POPULATION SIZE*

*In collaboration with Allan Strong*

## Objectives

- Explore how allele frequencies drift over time with stable populations of different sizes.
- Explore how allele frequencies drift over time when population sizes fluctuate.
- Calculate and interpret the effective population size of the population.

***Suggested Preliminary Exercises: Hardy-Weinberg Equilibrium; Genetic Drift***

## INTRODUCTION

The Hardy-Weinberg principle states that when populations are infinitely large, mate randomly, and experience no selection, mutation, or gene flow, both the allele and genotype frequencies can be predicted for the next generation. From a genetic perspective, infinitely large Hardy-Weinberg populations are considered "ideal" populations. That is, the number of males and females are equal, mating occurs randomly, all individuals contribute more or less equally to the next generation, and population size is large and does not vary over time. Thus, in a population with $N$ number of breeding individuals, each parent has a $1/N$ probability of producing a gamete that will be incorporated into future offspring.

But most, if not all, populations violate at least some of these assumptions. Population numbers fluctuate over time, have unequal sex ratios, or have mating systems where only a few dominant individuals breed, or disperse in such a way that not all individuals contribute equally to the next generation's genetic makeup. In other words, all of these "violations" can influence the way gametes are passed down to future generations.

How can we characterize populations that are *not* ideal? It is useful to directly compare the actual censused population size, $N_t$, to its **effective population size**, $\boldsymbol{N_e}$. The effective population size tells you how large the observed population is based on its genetic behavior. Because all populations have a finite size, they will experience some degree of genetic drift and inbreeding, even if the population is ideal in every other sense. The degree of drift and inbreeding in an ideal population with a finite size can be used as a baseline to which other, nonideal populations can be compared. You might recall from the preceding exercise that genetic drift is the change in allele frequency over generations that occurs because, by

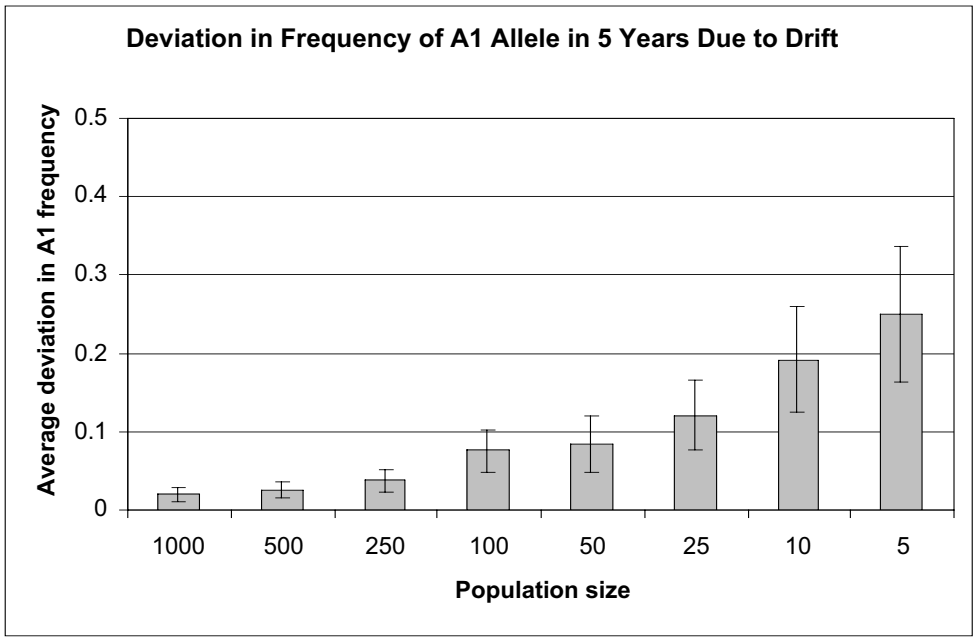


**Figure 1** In all cases, the starting frequency of the $A_1$ allele = 0.5. After 5 generations, the deviation in the allele frequency from 0.5 was recorded. You can see that small populations experience a significant amount of drift (change in allele frequency due to sampling error) compared to larger populations.

chance, alleles are not passed down to subsequent generations as predicted by Hardy-Weinberg. The smaller the population, the more drift occurs and the more likely alleles will become fixed. Figure 1 shows how much drift occurs over 5 generations in populations ranging in size from 1000 down to 5 individuals.

The concept of effective population relates directly to the concepts of genetic drift and inbreeding (Wright 1931). The effective size of a population, $N_e$, is the number of individuals that will contribute genes equally to the next generation. For example, suppose we count 270 turtles in a population (the censused population), and would like to know how those 270 turtles "behave" from a genetic standpoint. The effective population size tells us that number. If $N_e$ for this population equals 50, that means that our turtle population ($N_t$ = 270) behaves or experiences changes in its genetic makeup like an "ideal" population of 50 individuals (that is, a population where mating is random, sex ratios are even, individuals contribute gametes equally to the next generation, and population size does not vary over time, but that nonetheless experiences drift and inbreeding because the population is not infinite).

Often $N_e$ is less than $N_t$, suggesting that many natural populations behave genetically like a smaller population. A fluctuation in population size from year to year is one way that effective population size is reduced in nature. For example, suppose a population consists of 1000 individuals in generation 1, 10 individuals in generation 2, and 1000 individuals in generation 3. Generation 2 is considered a "bottleneck" generation for the population because only a handful of individuals actually survived through that period. Although we can count 1000 individuals in generation 3, the effective population size will be less than 1000 because the bottleneck has made the 1000 individuals in generation 3 more genetically related than the 1000 individuals in generation 1. In fact, this population will behave genetically more like an "ideal" population of 29 individuals ($N_e$ = 29). Therefore, the number of individuals contributing genetically to the next generation is less than the actual population size.

You may ask, "How did we arrive at the number 29 in the above example?" The number 29 is the **harmonic mean** of the numbers 1000, 10, and 1000, or the reciprocal of the

**average of the reciprocals** of these three numbers. In other words, one way of calculating $N_e$ is to compute the harmonic mean (see Crow and Kimura 1970 for greater detail). By using reciprocals to compute the harmonic mean, small numbers have a much greater effect than larger numbers. If $a = 10$ and $b = 2000$, then $a$ has much more influence on the harmonic mean than $b$ because 1/10 is much greater than 1/2000. Conceptually, this is exactly why computations of $N_e$ are based on harmonic means: The importance of inbreeding and genetic drift is much greater when the population is small than when it is large, so the smaller population numbers should be emphasized in any computation of $N_e$.

The **harmonic mean,** $N_e$, for populations that fluctuate in number can be calculated as

$$\frac{1}{N_e} = \frac{1}{t} \times \left(\frac{1}{N_1} + \frac{1}{N_2} + \ldots + \frac{1}{N_t}\right)$$

where $t$ is the number of years under consideration, and $N_1$, $N_2$, …, $N_t$ are the censused population sizes over time.

To be clear, let's walk through an example. Suppose we censused a population for 6 consecutive years, and counted 1000, 5, 5, 1000, 5, and 1000 individuals over time. The effective population size, $N_e$, is equal to the harmonic mean of 1000, 5, 5, 1000, 5, and 1000, and is calculated as

$$\frac{1}{N_e} = \frac{1}{t} \times \left(\frac{1}{N_1} + \frac{1}{N_2} + \ldots + \frac{1}{N_t}\right)$$

$$\frac{1}{N_e} = \frac{1}{6} \times \left(\frac{1}{1000} + \frac{1}{5} + \frac{1}{5} + \frac{1}{1000} + \frac{1}{5} + \frac{1}{1000}\right) = 10$$

$$\frac{1}{N_e} = 0.167 \times 0.603$$

$$N_e = \frac{1}{0.167 \times 0.603} = 10$$

This means that although we can count 1000 individuals in year 6, *genetically* the population is behaving like an ideal population of size 10.

In addition to fluctuating population size, effective population sizes are affected by **sex ratio, dispersal distances,** and **variation in offspring produced per female.** It's fairly straightforward to understand how mating systems and sex ratio can affect $N_e$. If a censused population of 100 individuals consists of only 2 female breeders and 10 male breeders, the gametes that are passed down to the future generation are strongly influenced by the genetic makeup of those breeders. Disperal distance affects $N_e$ because it determines how close or far siblings establish breeding sites from each other, which in turn affects the probability of mating with relatives. And variation in the number of offspring produced affects $N_e$ by altering which genes are incorporated into the next generation. For example, all females may breed in a given year, but if one or two females have "boom" years (reproduce a lot) while others have "bust" years, the variance in reproductive output is high. Obviously, these females do not contribute gametes equally to the next generation. It is beyond the scope of this exercise to discuss all of these factors (see Crow and Kimura 1970), but you should be aware that the effective size of natural populations is influenced in a variety of ways.

## PROCEDURES

The derivations for the various effective population size formulae are complicated, and therefore this exercise is devoted less to the math and more to explaining the genetic behavior of populations conceptually. In this exercise, we will simulate the effects of changes in gene frequencies for a population over the course of 6 generations. The first part of the exercise focuses on how much genetic drift occurs in populations with a

constant size. In each generation, the genotypes of individuals will be drawn according the Hardy-Weinberg theory, based on the genetic makeup of the parents in the preceding generation. We will assume that generations do not overlap and that individuals can self-fertilize—that is, the same parent can contribute both egg and sperm to produce an offspring. We will then allow populations to fluctuate so that you can observe the how much drift occurs when population sizes change over time. Additionally, we will construct a simple model to examine graphically the relationship between $N_t$ and $N_e$ over 6 generations. This part of the exercise will enable us to evaluate the effect of bottlenecks in $N_t$ on the effective populations size.

As always, save your work frequently to disk.

## INSTRUCTIONS

## ANNOTATION

***A. Set up the model population.***

1. Open a new spreadsheet and set up column headings as shown in Figure 2.

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ***Effective Population Size Simulation*** | | | | | | | | | | | | |
| 2 | | | **POPULATION SIZE** | | | | | | | | | | |
| 3 | | | **Gen. 1** | Parents | **Gen. 2** | Parents | **Gen. 3** | Parents | **Gen. 4** | Parents | **Gen. 5** | Parents | **Final** |
| 4 | **Allele freq.** | **Initial** | 1000 | | 1000 | | 1000 | | 1000 | | 1000 | | 1000 |
| 5 | **A1** | 0.5 | | | | | | | | | | | |
| 6 | **A2** | 0.5 | | | | | | | | | | | |

**Figure 2**

2. Enter 0.5 in cells B5 and B6.

We'll consider a population whose initial allele frequencies are $p$ = frequency of the $A_1$ allele = 0.5 and $q$ = frequency of the $A_2$ allele = 0.5. Remember that $p + q$ must equal 1 for loci that have only two alleles.

3. Enter the number 1000 in cells C4, E4, G4, I4, K4, and M4.

The cells C4, E4, G4, I4, K4, and M4 give the population size over generations. The final generation is given in cell M4. To begin, our population will have a constant size of $N_t$ = 1000. Later in the exercise we will vary these numbers. Shade these cells to remind you that they can be directly manipulated in the exercise.

4. In cells D4, enter the formula **=2*E4**. Enter analogous formulae into cells F4, H4, J4, and L4.

Cell D4 "controls" the *maximum* number of individuals from generation 1 that will survive and potentially produce offspring in generation 2. For example, generation 2 will consist of 1000 individuals, so up to 2000 randomly selected parents from generation 1 will produce them (i.e., 2000 gametes will be passed down from generation 1 to generation 2, and all 1000 individuals in generation 1 potentially contribute to the next generation's gene pool). If generation 2 consisted of only 10 individuals, we would let only 20 randomly selected parents potentially produce them (the first 20 individuals listed in the spreadsheet). If generation 2 consisted of 4000 individuals (for example), then all of the individuals in generation 1 would potentially produce offspring. Cell F4 "controls" the number of individuals from generation 2 that will contribute offspring to generation 3, etc.

By copying the D4 formula over to cells F4, H4, J4, and L4, the maximum number of parents will be determined by the population size in the next generation. Your formulae in those cells should be:

- F4 **=2*G4**
- H4 **=2*I4**
- J4 **=2*K4**
- L4 **=2*M4**

5. Save your work.

6. Set up new headings as shown in Figure 3.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | Genotype #'s | | | | | | | | | | | |
| 9 | A1A1 | | | | | | | | | | | |
| 10 | A1A2 | | | | | | | | | | | |
| 11 | A2A2 | | | | | | | | | | | |
| 12 | SUM | | | | | | | | | | | |
| 13 | | Individual | Gen. 1 | Parents | Gen. 2 | Parents | Gen. 3 | Parents | Gen. 4 | Parents | Gen. 5 | Parents |

**Figure 3**

7. Set up a linear series from 1 to 1000 in cells B14–B1013.

Enter **1** in cell B14.
Enter **=1+B14** in cell B15. Copy this formula down to cell B1013.
We will simulate the population dynamics over 6 generations. For any generation, the maximum population size can be 1000 (assuming the environment's carrying capacity will support 1000 individuals).

8. In cells C14–C1013, enter a formula to assign a genotype to individual 1 in generation 1 based on the frequencies given in cells B5–B6.

In cell C14 enter the formula **=IF(B14<=$C$4,IF(RAND()<$B$5,$A$5,$A$6) &IF(RAND()<$B$5,$A$5,$A$6),"")**. Copy the formula down to cell C1013

Use the **IF** function as you did in the Hardy-Weinberg exercise, with one IF function nested within another to control the population size according to the value in cell B14. Remember that the IF formula returns one value if a condition you specify is true, and another value if the condition you specify is false.

The first part of the formula in cell C14 tells the spreadsheet to determine if cell B14 is less than or equal to (<=) the value in cell C4. If so, carry out the function **IF(RAND()<$B$5,$A$5,$A$6)&IF(RAND()<$B$5,$A$5,$A$6)** to assign a genotype to the individual. If cell B14 is greater than the value in cell C14, return a double quote mark, **""** (which will return as a blank cell). This portion of the formula controls the population size. The genotype assignment is the same as you did in the Hardy-Weinberg exercise: The function tells the program to choose a random number between 0 and 1 (the **RAND()** part of the formula). If that random number is less than the value designated in cell B5 (the frequency of the $A_1$ allele), then assign it an allele of $A_1$; otherwise, assign it a value of $A_2$. Since all individuals have two alleles for a given locus, the formula is repeated again and genotype is generated by joining the two alleles with an **&** symbol. Once you've obtained genotypes for individual 1, copy this formula down to cell C1013 to obtain genotypes for all 1000 individuals in the population in generation 1.

9. Enter a formula in cells C9–C11 to count the number of individuals of each genotype in generation 1.

In cell C9 enter the formula **=COUNTIF(C14:C1013,"A1A1")**.
In cell D9 enter the formula **=COUNTIF(C14:C1013,"A1A2")+COUNTIF(C14:C1013, "A2A1")**.
In cell E9 enter the formula **=COUNTIF(C14:C1013,"A2A2")**.
You are using the **COUNTIF** function to count the various genotypes in generation 1. Don't forget that heterozygotes can be either $A_1A_2$ or $A_2A_1$. Double-check your results in the next step.

10. Sum the genotypes in Generation 1 in cell C12.

In cell C12 enter the formula **=SUM(C9:C11)**. Your result should be 1000.

11. Enter formulae in cells C5 and C6 to compute the actual allele frequencies in generation 1.

In cell C5 enter the formula **=(2*C9+C10)/(2*C12)**.
In cell C6 enter the formula **=1-C5** or **=(2*C11+C10)/(2*C12)**.
Remember from the Hardy-Weinberg exercise that you can compute the allele frequencies easily if you know the genotype frequencies. The equations are freq($A_1$) = $p$ = ($2N_{A1A1} + N_{A1A2}$) / $2N$, where $N$ is the total number of individuals in the population. The frequency of the $A_2$ allele can be computed either by subtraction (= 1 – $p$), or by freq($A_2$) = $q$ = ($2N_{A2A2} + N_{A1A2}$) / $2N$.

12. In cells D14–D1013, enter a formula to select the parents that can potentially produce offspring in the next generation.

In cell D14 enter the formula **=IF(B14<=$D$4,C14,"")**. Copy this formula down to cell D1013.
The formula in cell D14 identifies the parents. The allele frequencies of this parental population will be used to assign genotypes to individuals in generation 2. If cell B14 (individual 1) is less than or equal to the maximum number of parents in generation 1, the program will return individual 1's genotype. Otherwise, it will return a blank cell (the double-quote marks).

13. Copy cells C5–C6 and C9–C12 across to cells L5–L6 and L9–L12.

14. In cells E14–L14, enter formulae for the remaining generations, and copy your formulae down to row 1013 of each column as you go. Save your work.

This action will allow you to obtain genotype numbers and allele frequencies of the parents in generation 1, as well as future generations and parents. The entries for future generations will not make sense until you have completed the next step.

Follow the examples from generation 1, but make sure you update the formulae appropriately. Pay attention to absolute and relative references, and make sure that the new generation is based on the allele frequencies of the parental generation preceding it. Double-check your formulae.

We used the following formulae:

- Cell E14 **=IF(B14<=$E$4,IF(RAND()<$D$5,$A$5,$A$6)&IF(RAND()<$D$5,$A$5,$A$6),"")**
- Cell F14 **=IF(B14<=$F$4,E14,"")**
- Cell G14 **=IF(B14<=$G$4,IF(RAND()<$F$5,$A$5,$A$6)&IF(RAND()<$F$5,$A$5,$A$6),"")**
- Cell H14 **=IF(B14<=$H$4,G14,"")**
- Cell I14 **=IF(B14<=$I$4,IF(RAND()<$H$5,$A$5,$A$6)&IF(RAND()<$H$5,$A$5,$A$6),"")**
- Cell J14 **=IF(B14<=$J$4,I14,"")**
- Cell K14 **=IF(B14<=$K$4,IF(RAND()<$J$5,$A$5,$A$6)&IF(RAND()<$J$5,$A$5,$A$6),"")**
- Cell L14 **=IF(B14<=$L$4,K14,"")**

Review your formulae and double-check your work. Make sure you understand the formulae (and model) before proceeding.

***B. Compute changes in $A_1$ due to genetic drift.***

1. In cell M5, compute the deviation in the $A_1$ allele as the difference between the initial frequency in cell B5 and the final frequency in cell L5.

In cell M5 enter the formula **=ABS(L5-B5)**. Enter a label for this value in cell N5 as shown in Figure 4.
This is simply the **absolute value** of the difference between the initial and final frequency of the $A_1$ allele. It merely quantifies how far the $A_1$ allele drifted—we don't care about which direction the allele drifted.

| | M | N |
|---|---|---|
| 3 | **Final** | |
| 4 | 1000 | |
| 5 | | <= deviation |

**Figure 4**

Remember that so far our population is ideal, except that it is finite—it consists of 1000 individuals over the generations. Any change in allele frequencies is due solely to genetic drift because the model does not include gene flow, natural selection, mutation, or nonrandom mating.

2. Press F9 to run a new simulation. What level of drift did the population experience?

You should see that the level of drift varies each time you press F9, the calculate key. This is because of the random way in which genotypes are assigned to individuals in each generation based on the Hardy-Weinberg principle. In order to "quantify" the level of drift, we will run 100 simulations, each time recording the deviation in frequency of the $A_1$ allele from the initial conditions. The average and standard deviation of these simulations will give a better indication (quantification) of the level of drift the population experienced after five generations and a constant population size of $N_t = 1000$.

3. Set up new headings as shown in Figure 5, except extend your trials to 100 (cell O103).

| | O | P | Q |
|---|---|---|---|
| 2 | | **Drift of A1** | |
| 3 | Trial | *N* = 1000 | *N* = 10 |
| 4 | 1 | | |
| 5 | 2 | | |
| 6 | 3 | | |
| 7 | 4 | | |
| 8 | 5 | | |

**Figure 5**

4. Develop a macro to track drift over 100 simulations – track your results in cells P4–P103.

Open the macro program and assign a shortcut key (refer to Exercise 2 for details on building macros). In Record mode, perform the following steps:

- Press F9 to obtain a new set of random numbers, and hence a new set of genotypes for the populations.
- Select cell M5, the change in frequency of the $A_1$ allele due to drift, then open Edit | Copy.
- Select cell P3, the column labeled "N = 1000".
- Open Edit | Find. In the dialog box, leave the Find What box empty, searching by columns and formulas, and then select Find Next and Close.
- Open Edit | Paste Special | Paste Values. Click OK.
- Open Tools | Macro | Stop Recording.

Now press your shortcut key until 100 simulations have been recorded.

5. In cell P104, enter a formula to compute the average deviation in the $A_1$ allele due to drift.

In cell P104, enter the formula **=AVERAGE(P4:P103)**.

6. In cell P105, compute the standard deviation of the 100 simulations.

In cell P105, enter the formula **=STDEV(P4:P103)**.

7. In cell P106, enter **=P105/2**.

For graphing purposes, we will divide the standard deviation by 2 so that when the standard error bars are added to our graph (next section), half of the line will be above the mean and half will be below it.

8. Change your population numbers so that each generation consists of 10 individuals, as in Figure 6.

| | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | **Gen. 1** | Parents | **Gen. 2** | Parents | **Gen. 3** | Parents | **Gen. 4** | Parents | **Gen. 5** | Parents | **Final** |
| 4 | 10 | | 10 | | 10 | | 10 | | 10 | | 10 |

**Figure 6**

Now we will compare drift for a fixed population size of $N_t = 10$.

9. In column Q, develop a new macro to record deviations in the $A_1$ allele for this population.

See Step 4.

10. Copy cells P104-P106 to cells Q104–Q106.

This will generate means and standard deviations for this population, whose size is fixed at 10 individuals across generations.

***C. Create graphs.***

1. Graph the average deviation of the $A_1$ allele due to drift for the population when $N = 1000$ versus $N = 10$.

Use the column graph option. Under the Series tab, select cells P3 and Q3 as $x$-axis labels. Your graph should resemble Figure 7.

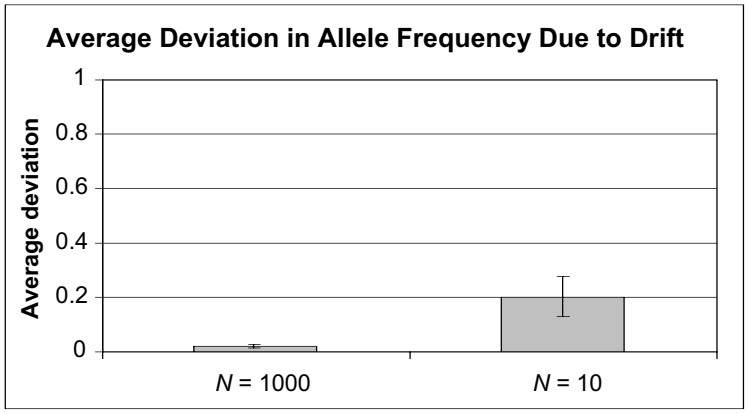


**Figure 7**

2. Add error bars to your graph.

To add error bars to your graph, click once somewhere in one of the columns in your graph. Go to Format | Selected Data Series. In the dialog box (Figure 8), select Y-Error Bars, then select the Display Both option for displaying error bars. Under Error Amount, select the Custom option. Select cells P106–Q106 in the + box, and repeat for the – box. Click OK and error bars will be added to your graph.

3. Save your work. We will interpret your model results and explore how fluctuating population size affects the level of drift in a population in the Questions section.


**Figure 8**

## QUESTIONS

1. Compare the drift in the $A_1$ allele for the population of $N = 1000$ (constant over time) and the population of $N = 10$ (constant over time). Which population shows a greater level of drift? Why?
2. When populations fluctuate, they "behave" like smaller populations that have a constant population in that they experience genetic drift in similar ways. Alter your spreadsheet so that the population size for generations is

- Generation 1 = 1000
- Generation 2 = 5
- Generation 3 = 5
- Generation 4 = 1000
- Generation 5 = 5
- Final generation = 1000.

The final generation consists of 1000 individuals, yet the effective population size, as computed with the formula is 10:

$$\frac{1}{N_e} = \frac{1}{6} \times \left( \frac{1}{1000} + \frac{1}{5} + \frac{1}{5} + \frac{1}{1000} + \frac{1}{5} + \frac{1}{1000} \right) = 10$$

This means that the fluctuating population will change in allele frequencies through drift in a way a constant population of size 10 will. Prove this to yourself by running a new macro (record the results in column R) and comparing your results to the constant, small population size. Graph your results.

3. Directly compute $N_e$ for your 6 generations. Set up the following new headings:

| | S | T | U | V | W | X |
|---|---|---|---|---|---|---|
| 2 | Generation | *Nt* | 1/*Nt* | Sum 1/*Nt* | *Ne* | HARMEAN |
| 3 | 1 | 1000 | 0.001 | 0.001 | 1000 | 1000 |
| 4 | 2 | 5 | | | | |
| 5 | 3 | 5 | | | | |
| 6 | 4 | 1000 | | | | |
| 7 | 5 | 5 | | | | |
| 8 | 6 | 1000 | | | | |

Enter formulae in cells T3–T8 to link population sizes given in cells C4, E4,…,M4. Enter a formula in cells U3–U8 to compute $1/N$. In cells V3–V8, enter formulae to track the sum of $1/N$ as more generations are considered. Finally, enter a formula in cell W3 to compute $N_e$. Refer back to the introduction for your computations. Graph how $N_e$ and $N_t$ change over time, and fully interpret your graph.

4. Explore the spreadsheet function **HARMEAN**, which computes the harmonic mean of a series of numbers directly in column X. For any given series of numbers, when is the harmonic mean the highest possible value? When is it the lowest possible value? For any given series of numbers, under what conditions is $N_e > N_t$? Explore you model by changing values of $N_t$, increasing and decreasing the variation in numbers over time. Pay attention to how $N_e$ is affected by bottlenecks both in the current generation and in subsequent generations.

## LITERATURE CITED

Crow, J. F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Lande, R., and G. F. Barrowclough. 1987. Effective population size, genetic variation, and their use in population management. *In* M. E. Soulé (ed.), *Viable Populations for Conservation*, pp. 87–123. Cambridge University Press, Cambridge.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97–159.