MATH MINUTES
diatom_sim.xls

Math Minute 4.2 How Do You Model Population Diversity?

After 11 days of diatom sampling, Rynearson and Armbrust were able to genotype 607 cells with 496 unique genotypes. What if they had been able to genotype 1,000 times as many cells? Would they have found 1,000 times as many unique genotypes? Probably not. With continued sampling, a growing percentage of cells would have had a previously sampled genotype. Of course, we could never collect and genotype every cell in the diatom population, so we must rely on sampling methods and mathematical tools to model population diversity. In this Math Minute, we explore how to use sample data to estimate the total number of unique genotypes in a population, and to predict the number of genotypes we should see once, twice, three times, and so on, in a sample of 607 cells.

Several different mathematical models are available for predicting the number of unique genotypes in a population based on the number of each genotype in a sample. Rynearson and Armbrust applied two of these models: Chao1 and abundance-based coverage estimators (ACE). These models are adapted from mark-release-recapture statistics used in field biology to estimate animal population sizes. To compute Chao1, square the number of genotypes observed only once, divide by two times the number of genotypes observed exactly twice, and add the number of different genotypes observed. Using the data from Figure 4.6b, we have

$$\text{Chaol} = \frac{427^2}{2 \times 47} + 496 = 2{,}436.$$

The ACE estimate is similar to Chao1, but includes terms for the number of genotypes observed *k* times, for *k* between 1 and 10, and is therefore a more complicated formula. We won't go through the details, but for Figure 4.6b, ACE estimates 2,747 different genotypes. From the Chao1 and ACE estimates, the investigators inferred that at least 2,400 genotypes were in the population, and they used a rough average of the two calculated values (2,550) in their subsequent statistical analyses.

Figure 4.6b shows the expected and observed numbers of genotypes that appear *k* times in our sample, for *k* between 1 and 7. The expected numbers are computed under the assumption that all 2,550 genotypes are equally abundant in the population. The equal-abundance assumption is not supported by the data, because the expected numbers are so different from the observed. Can we know how many genotypes we should see exactly once, twice, etc.?

The investigators chose a Poisson distribution to model the number of genotypes seen each number (1–7) of times. As you might have learned in a probability or statistics class, the Poisson distribution is a good model for frequencies, i.e. how many times a single event occurs. For example, if we were just trying to model how many times genotype #42 occurs in the sample of 607 cells, the Poisson distribution would be a good method. But we are trying to model frequencies of frequencies—that is, how many genotypes were sampled each number of times (e.g.,13 genotypes were sampled 3 times)—so there is no single "event" to count. A better model for the expected values in Figure 4.6b can be constructed using more advanced statistical methods. Rather than discuss this theoretical model, we will determine the expected values using a simulation. Simulations are good tools for modeling complex situations like this diatom-sampling data set, and simulations can also be used to validate theoretical models.

A simulation uses random numbers to represent objects or events. In this simulation (download diatom_sim.xls from the web site), each cell type is represented by a random number between 1 and 2,550, the assumed number of genotypes in the entire diatom population. The list of 607 randomly generated numbers represents the sampling and genotyping processes: pull out a cell, genotype it, record the genotype, and repeat this process a total of 607 times. These 607 genotype observations represent one run of the experiment. Once the list of 607 numbers (i.e., genotypes) is generated, you just need to count how many times each genotype occurred, then count how many genotypes occurred once, twice, three times, and so on.

Because each simulated list of 607 cell genotypes is random, you will get slightly different results every time you repeat the experiment. The idea of a simulation is that you can repeat the experiment many times in silico, and average the results to get a good estimate of what to expect in any given run of a nonsimulated, biological experiment. In the Math Minute Discovery Questions, you will use the Excel file diatom_sim.xls to simulate the diatom-sampling experiment and determine the expected numbers of genotypes you should see each number of times. Because the Poisson model used by the investigators was not the best model for the genotype frequency data, your expected values will not agree with those in Figure 4.6b. However, as you will see, the difference between expected and observed values for genotypes occurring four or more times is even greater than under the Poisson model. Therefore, the investigators' conclusion that all 2,550 genotypes are not equally abundant is still supported.

MATH MINUTE DISCOVERY QUESTIONS

1. Run the diatom-sampling simulation in diatom_sim.xls, repeating the experiment 10 times, and recording the average number of genotypes occurring *k* times, for *k* between 1 and 5. Record your results when you run the simulation 3 times for 10 repetitions each time.
2. Repeat the process from Math Minute Discovery Question 1 another 3 times, but increase the repetitions to 100 for each of the 3 iterations. Compare your answers to those for 10 repetitions. Is it better to use 10 or 100 repetitions? Explain your answer using the data from the simulations.
3. Continue to run the diatom-sampling simulation in repetitions of 100 until at least one genotype occurs more than 5 times (i.e., the "Average" row does not have zeroes in the two columns labeled 6 and 7). How many total repetitions did it take? Compare your answer to the answers of three other people (or repeat the process three more times). Use these four numbers to estimate the probability of observing a genotype more than five times in a single run of the sampling experiment.

Math Minute 4.2 How Do You Model Population Diversity?

1. Run the diatom sampling simulation in diatom_sim.xls, repeating the experiment 10 times, and recording the average number of genotypes occurring *k* times, for *k* between 1 and 5. Record your results when you run the simulation 3 times for 10 repetitions each time.

Sample results of three simulations, with 10 repetitions each, are shown below. The numbers should vary from time to time and student to student.

1	2	3	4	5
482.4	55.9	4	0.2	0
475.5	58.8	4.5	0.1	0
472.6	58.1	5.1	0.6	0.1

2. Repeat the process from Math Minute Discovery Question 1 another 3 times, but increase the repetitions to 100 for each of the 3 iterations. Compare your answers to those for 10 repetitions. Is it better to use 10 or 100 repetitions? Explain your answer using the data from the simulations.

Sample results of three simulations, with 100 repetitions each, is shown below. The numbers should vary from time to time and student to student.

1	2	3	4	5
475.98	58.07	4.53	0.31	0.01
478.77	56.71	4.59	0.26	0
477.61	57.29	4.61	0.22	0.02

It is better to use 100 repetitions, because there is less variation in the number of genotypes that occur each number of times. This is an example of the law of averages at work: as the number of repetitions increases, the average converges to the true mean.

3. Run the diatom sampling simulation with 500 repetitions. Based on your simulation, how many genotypes do you expect to observe 4 times and 5 times? Compare your results to the expected numbers in Figure 4.6b. How does this difference affect your interpretation of the results of this study?

Expected number of genotypes observed 4 and 5 times should be approximately 0.267 and 0.013, respectively. These results are significantly smaller than the expected numbers in Figure 4.6b, so there is an even bigger difference in observed and expected than shown in the figure. This larger difference in observed and expected lends greater strength to the hypothesis that all genotypes are not equally abundant in the population.