Math Minute 8.1 **How Do You Know if You Have Sampled Enough Cells?**

A hidden assumption in the bar code study is that the proportion of each strain in a sample is the same as the proportion of the strain in the entire population at the time the sample is taken. If this assumption does not hold, the growth profiles of the 558 strains cannot be accurately assessed. The validity of the assumption depends on how large the

sample is, compared to the size of the population. For example, suppose a population of 10^{10} cells contains an equal number of cells from each of 558 strains. If 10^9 cells (10%) are sampled from the population, it is more likely that the sample contains an approximately equal number of cells from each strain than if only 10^8 cells (1%) are sampled. If one strain were to constitute only 1% of the population, a very small sample might miss the strain completely, or get too few cells from the strain to be detected on a DNA microarray.

You can calculate the probability of a particular sampling outcome (i.e., getting a particular number of cells from each strain in a sample) using the multivariate **hypergeometric** frequency function, a formula involving the sample size and number of cells from each strain in the population. Specifically, suppose a population contains N cells from M different strains, with n_1 cells from strain 1, n_2 cells from strain 2, . . . and n_M cells from strain M. You can compute the probability that a random sample contains k_1 cells from strain 1, k_2 cells from strain 2, . . . and k_M cells from strain M, with the hypergeometric formula:

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}\binom{n_3}{k_3}\cdots\binom{n_M}{k_M}}{\binom{N}{K}}$$

where K is the sample size (i.e., $K = k_1 + k_2 + \ldots + k_M$). For example, if a population contains 100 cells from each of 3 different strains (a total of 300 cells), the probability that a sample of 60 cells contains exactly 20 cells from each strain is

$$\frac{\binom{100}{20}\binom{100}{20}\binom{100}{20}}{\binom{300}{60}}$$

In the hypergeometric formula, each pair of numbers in parentheses is a **binomial coefficient**

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

(read "n choose k"), which represents how many distinct sets of k cells of a particular strain can be chosen from the n cells of that strain in the population. For example, with $n_1 = 100$ and $k_1 = 20$,

$$\binom{n_1}{k_1} = \binom{100}{20} = \frac{100!}{20!(100-20)!} \approx 5.36 \times 10^{20}$$

Many calculators, as well as various mathematical and statistical software programs, have a binomial coefficient function that can save you a lot of computation.

In a sample of 60 cells from a population of 300 cells, the ideal sample would contain 20 cells from each strain, in perfect agreement with the population proportions of 1/3 for each strain. Evaluating the binomial coefficients in the preceding formula results in a probability of 0.017 of getting an ideal sample, that is, one that contains exactly 20 cells from each of the three strains. However, for practical purposes, we are satisfied with getting *close* to 20 cells from each strain. Suppose we are willing to accept a deviation of up to three cells from the ideal number (20 ± 3) for each strain, corresponding to a deviation of $3/60 = 1/20 = 0.05$ from the ideal proportion ($1/3 \pm .05$). There are 37 different sampling outcomes that satisfy this criterion. (Can you identify all 37?)

You can find the probability of getting one of these 37 outcomes by computing the probability of each outcome (using the hypergeometric formula) and adding the 37 probabilities. The result of these calculations—that is, the probability that the sample proportions are all within 0.05 of the population proportions—is approximately 0.47.

Table MM8.1 Probability that sample proportions are within a specified deviation from 1/3, for sample sizes of 60,120, and 180 cells.

	Probability		
Deviation	**Sample 60 cells**	**Sample 120 cells**	**Sample 180 cells**
0.025	0.113	0.345	0.502
0.05	0.470	0.766	0.954
0.075	0.649	0.954	0.998
0.1	0.887	0.995	0.999991
0.125	0.945	0.9997	≈1

In other words, we have a 53% chance of getting a sample that *differs* from the ideal by more than 3 cells in one or more strains.

To improve our chances of getting a good sample, we must sample more cells. Alternatively, we could relax our maximum deviation criterion, accepting a deviation of up to 5 cells (20 ± 5), for example. The probability that all strains are represented within a given deviation from 1/3 is shown in Table MM8.1 for three different sample sizes. The deviation is given as a proportion of the sample size; the deviation in number of cells is different for each sample size.

This table of probabilities shows that we should sample at least 180 cells from our population of 300 cells to be 95.4% certain that the sample proportions will be $1/3 \pm 0.05$ for all three strains. However, if we were willing to accept errors as large as 0.125 (i.e., sample proportions ranging from 0.208 to 0.458), a sample of 60 cells would probably do.

Now suppose we have a population containing 200 cells from strain 1; 97 cells from strain 2; and 3 cells from strain 3. In other words, the population is the same size as in our previous example (300 cells), but strain 3 constitutes only 1% of the population. What is the probability that a sample of 60 cells from this population contains at least one cell from strain 3? There are 61 sampling outcomes that have no cells from strain 3. By using the hypergeometric formula, you can sum the probabilities of these 61 outcomes to find that the probability of completely missing strain 3 is 0.511. Therefore, the probability that strain 3 is present in the sample of 60 cells is $1 - 0.511 = 0.489$. Once again, sampling more cells improves our chances of getting a good sample—in this case, one in which a cell from strain 3 is present. Specifically, the probability that strain 3 is present is 0.785 when 120 cells are sampled, and 0.937 when 180 cells are sampled.

For very large samples, it is difficult to compute probabilities with the hypergeometric frequency function, and they are often approximated using the **normal distribution** (Math Minute 11.1). Whether exact or approximate, these probabilities show investigators that growth trends such as those in Figure 8.8 are not merely artifacts of the sampling process.