EXAMPLE A Suppose that we wish to find the expectation of a binomial random variable, Y. From the binomial frequency function,

$$E(Y) = \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k}$$

It is not immediately obvious how to evaluate this sum. We can, however, represent Y as the sum of Bernoulli random variables, X_i, which equal 1 or 0 depending on whether there is success or failure on the ith trial,

$$Y = \sum_{i=1}^{n} X_i$$

Because $E(X_i) = 0 \times (1-p) + 1 \times p = p$, it follows immediately that $E(Y) = np$.

An application of the binomial distribution and its expectation occurs in "shotgun sequencing" in genomics, a method of trying to figure out the sequence of letters that make up a long segment of DNA. It is technically too difficult to sequence the entire segment at once if it is very long. The basic idea of shotgun sequencing is to chop the DNA randomly into many small fragments, sequence each fragment, and then somehow assemble the fragments into one long "contig." The hope is that if there are many fragments, their overlaps can be used to assemble the contig.

Suppose, then, that the length of the DNA sequence is G and that there are N fragments each of length L. G might be at least 100,000 and L about 500. Assume that the left end of each fragment is equally likely to be at positions $1, 2, \ldots, G - L + 1$. What is the probability that a particular location $x \in \{L, L+1, \ldots, G\}$ is covered by at least one fragment? How many fragments are expected to cover a particular location? (The positions $\{1, 2, \ldots, L-1\}$ are not included in this discussion because the boundary effect makes them a little different; for example, the only fragment that covers position 1 has its left end at position 1.) To answer these questions, first consider a single fragment. The chance that it covers x equals the chance that its left end is in one of the L locations $\{x - L + 1, x - L + 2, \ldots, x\}$, and because the location of the left end is uniform, this probability is

$$p = \frac{L}{G - L + 1} \approx \frac{L}{G}$$

where the approximation holds because $G \gg L$. Thus, the distribution of W, the number of fragments that cover a particular location, is binomial with parameters N and p.

From the binomial probability formula, the chance of coverage is

$$P(W > 0) = 1 - P(W = 0) = 1 - \left(1 - \frac{L}{G}\right)^N$$

Since N is large and p is small, the distribution of W is nearly Poisson with parameter $\lambda = Np = NL/G$. From the Poisson probability formula, $P(W = 0) \approx e^{-NL/G}$, so the probability that a particular location is covered is approximately $1 - e^{-NL/G}$. Observe that NL is the total length of all the fragments; the ratio NL/G is called the *coverage.* Calculations of this kind are thus useful in deciding how many fragments to use. If the coverage is 8, for example, the chance that a site is covered is .9997.