Overlap of fragments is important when trying to assemble them. Since W is a binomial random variable, the expected number of fragments that cover a given site is $Np = NL/G$, precisely the coverage.

We can also now answer this closely related question: How many sites do we expect to be entirely missed? We will calculate this using indicator random variables: let I_x equal 1 if site x is missed and 0 elsewhere. Then

$$E(I_x) = 1 \times P(I_x = 1) + 0 \times P(I_x = 0) = e^{-NL/G}.$$

The number of sites that are not covered is

$$V = \sum_{x=1}^{G} I_x$$

and from the linearity of expectation

$$E(V) = \sum_{x=1}^{G} E(I_x) \approx Ge^{-NL/G}.$$

The length of the human genome is approximately $G = 3 \times 10^9$, so with eight times coverage, we would expect about a million sites to be missed. ■

EXAMPLE **B** *Coupon Collection*

Suppose that you collect coupons, that there are n distinct types of coupons, and that on each trial you are equally likely to get a coupon of any of the types. How many trials would you expect to go through until you had a complete set of coupons? (This might be a model for collecting baseball cards or for certain grocery store promotions.)

The solution of this problem is greatly simplified by representing the number of trials as a sum. Let X_1 be the number of trials up to and including the trial on which the first coupon is collected: $X_1 = 1$. Let X_2 be the number of trials from that point up to and including the trial on which the next coupon different from the first is obtained; let X_3 be the number of trials from that point up to and including the trial on which the third distinct coupon is collected; and so on, up to X_n. Then the total number of trials, X, is the sum of the $X_i,\ i = 1, 2, \ldots, n$.

We now find the distribution of X_r. At this point, $r - 1$ of n coupons have been collected, so on each trial the probability of success is $(n - r + 1)/n$. Therefore, X_r is a geometric random variable, with $E(X_r) = n/(n - r + 1)$. (See Example B of Section 4.1.) Thus,

$$\begin{aligned} E(X) &= \sum_{r=1}^{n} E(X_r) \\ &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} \\ &= n \sum_{r=1}^{n} \frac{1}{r} \end{aligned}$$

For example, if there are 10 types of coupons, the expected number of trials necessary to obtain at least one of each kind is 29.3.