Chapter 5

Probabilistic analysis of sequencing methods

Modern molecular biology makes it possible to sequence DNA quickly and easily and and computer science and mathematical analysis make it possible to scrutinize huge data sets of sequences for biological information. Methods for obtaining and then interpreting biological sequences have developed at an amazing pace. In Spring of 2001, the Human Genome Project and, independently, Celera Inc. under the direction of Craig Ventner, announced working drafts of the entire human genome, about 3 billion base pairs long. The Human Genome Project cost US taxpayers 2.7 billion dollars. The rough cost of sequencing one genome at this time, excluding the cost of the the machinery and software was on the order of $100,000,000. By the end of 2013 the cost per genome using commercially available machines was well under $10,000. In January of 2014, the company Illumina announced that running 10 of its latest machines in parallel makes it possible to sequence a human genome for about $1,000—this figure excludes the hefty, one-time cost of the 10 machines. Because of the availability of relatively cheap sequencing, sequence analysis is now a standard working tool of the biological scientist. Sequenced DNA from chromosomes and mitochondria provides basic data for studying problems in genetics, evolutionary biology, and medicine, and for the discovery of new genes and proteins.

This chapter is a modest introduction to some quantitative models for sequence analysis. We first study *shotgun sequencing*. This is a commonly used method for reconstructing long sequences—an entire chromosome or genome, for example—from the relatively short segments that can be directly sequenced by current technology. Then we study *restriction enzyme digests*, which chemically fragment DNA with special enzymes. Digests are used for sequencing, localizing genes, genetic mapping, and other applications. In both cases, we will study average, physical properties of the techniques, not the biological information they produce. The models in this chapter are basic, but have several virtues. They are simple, they are useful, and they introduce tools that are used widely in probabilistic modeling.

An important theme of this chapter is approximation. The models are approximate; they use continuous random variables for processes that are really discrete and make simplifying assumptions on the probability distributions. The main formulas are also approximate; their derivations ignore small complicating effects, trading complete accuracy for simplicity. It is important to learn how to use such simplifications and approximations in applied modeling, and the examples of this chapter are good paradigms to study.

A second theme is the application of the Poisson random variable. In particular, we will introduce and apply a continuous-time random process called the Poisson process, based on the Poisson distribution. There is a relation with approximation here. The Poisson random variable approximates a binomial random variable with parameters n and p, when n is large and p is small, and that is how it arises in applications to restriction enzyme digests.

We repeatedly use an approximation that should be familiar to you from beginning calculus. Recall that for any real number x,

$$\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n = e^{-x}. \tag{5.1}$$

An informal statement of this fact is:

$$\left(1-\frac{x}{n}\right)^n \approx e^{-x} \quad \text{for large } n, \tag{5.2}$$

where, as usual, "$\approx$" means "is approximately equal to." Of course this is not a precise statement. For what range of x and n is the approximation good? For $x > 0$, it can be shown that

$$0 < e^{-x} - \left(1-\frac{x}{n}\right)^n < \min\{e^{-x}, \frac{x}{n}\} \qquad \text{when } x/n < 1/2. \tag{5.3}$$

Hence the error of the approximation is at most of the order of x/n. A somewhat unusual looking approximation is obtained from (5.3) if we set $x = \delta n$: for "small" δ,

$$(1-\delta)^n \approx e^{-n\delta} \tag{5.4}$$

The inequality of (5.3) shows that the error of this approximation is no more than $\min\{e^{-n\delta}, \delta\}$, no matter what n is.

Exercise 5.0.1. To get a sense of the approximation in (5.3), compute $e^{-x} - (1-\frac{x}{n})^n$ and $(1-x)^n/e^{-x}$ for $n = 50$ and $n = 100$ when $x = 1$ and $x = 5$.

Exercise 5.0.2. a) Define $R(y)$ so that $\ln(1-y) = -y + R(y)$. Use Taylor's remainder theorem to show that $\dfrac{y}{2(1-y)^2} < R(y) < 0$.

b) Using $e^{-x} - (1-x/n)^n = e^{-x}[1 - e^{x+n\ln(1-x/n)}]$ and the result of a), show that $0 < e^{-x} - (1-x/n)^n < xe^{-x}[x/n][1/2(1-(x/n))^2]$. Hint: $1 - e^{-s} < s$ for $s > 0$.

c) Use the result of b) to show (5.3). Hint: Find the maximum of xe^{-x} for $x > 0$.

5.1 Shotgun sequencing

5.1.1 The method and the coverage problem

Celera's draft of the human genome was announced in the paper, Ventner, et al, "The Sequence of the Human Genome," *Science* (2001) Vol. 291, 1304-1351. The abstract of the paper begins: "A 2.91 billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method." Lest the gentle reader fear entering a modern genetics laboratory, rest assured that shotgun sequencing does not require firing rifles. Here, the adjective *shotgun* connotes "covering a wide area in an irregularly effective manner...." (The Random House Dictionary of the English Language, second edition). Sequencing is ultimately based on automated chemical methods that work directly only on relatively short segments. Sequencing a segment is called a *read*. The segment length, or read length, that a lab machine can handle depends on the method, but is typically on the order of a few hundred to a thousand base pairs. (A new technique based on threading DNA through nanometer sized holes in a membrane is being developed and holds the promise of much greater read lengths. See *Science*, 21 February 2014, page 839.) Whatever the sequencing technology, biologists want complete DNA sequences of segments, such as the DNA of a whole chromosome, many orders of magnitude longer. Shotgun sequencing is a strategy for tackling a long segment by randomly generating smaller fragments of it that can be directly sequenced.

The method is simple, almost obvious, in principle. We will explain it, for the problem of sequencing one long, single-stranded DNA segment, which we will call S. First make multiple copies of S, and fragment them into pieces small enough to be sequenced. This can be done by a mechanical process that shears the DNA copies randomly. Figure 5.1 schematically illustrates several copies (clones) of a segment which have each been divided randomly into the fragments defined by the short vertical lines. In the figure, the fragments are shown attached in their proper order on the segment copy they belong to, but in the laboratory procedure they are all detached from one another and the fragments of all copies are mixed in a common pool.

In Figure 5.1, two fragments F_1 and F_2 from different copies of S are labeled. Notice that they occupy overlapping subintervals of S. In the part that overlaps, comprising a piece of the right end of F_1 and piece of the left end of F_2, they must share a common sequence of basis. This is the observation on which shotgun sequencing is based. If we were given only the sequences in F_1 and F_2, without being told where they came from in S, we could deduce that they overlapped from the fact that they share a common sequence at their ends. Likewise, if two fragments shared no such common sequence, we would know that they came from disjoint parts of S. Of course, we are exaggerating a bit. It might be that two disjoint fragments, such as F_1 and F_3, by chance share a common sequence at different ends.

For example, suppose that F_1 ended in A and F_3 began with A; there is certainly a good chance that this can occur and it is a case of a shared common sequence, even if this sequence only has length one. We should really say that when two fragments share a common sequence at opposing ends there is a probability that they overlap, the probability being higher the longer the common sequence is.

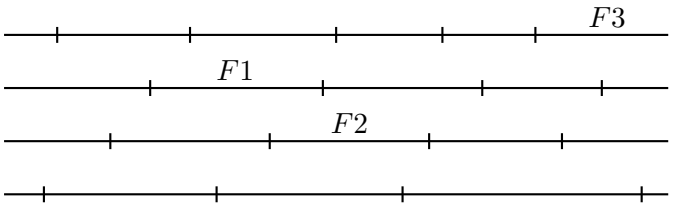

Figure 5.1. Copies of S differently fragmented.

Return now to our pool of separated, mixed fragments. The idea of the shotgun method is to randomly select some number, N, of the fragments, to sequence them, and, by searching through these sequences for common overlapping subsequences, to connect them into longer, fully sequenced pieces of S. As a first approximation, assume that any common subsequence at the ends of two fragments means a true overlap, so that no false connections are made. The next figure shows a hypothetical outcome of such an analysis. The line segment at the bottom represents the original piece S of DNA. Above S, we have placed the $N = 14$ segments that were sequenced, positioning them in their proper (but *a priori* unknown) positions along S. Above them, we have represented the longer contiguous portions $C1$, $C2$, and $C3$ of S that have been fully sequenced by connecting fragments with common ends.

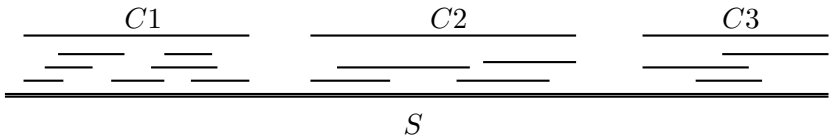

Figure 5.2: S with sequenced fragments, contigs, and gaps.

The segments $C1$, $C2$ and $C3$ are called **contigs**, from the word *contiguous*. They are contiguous stretches of linked, sequenced fragments which cannot be extended using any of the other N sequenced fragments. In between the sequenced contigs are gaps, regions which, by chance, were not covered by any of the N sequenced fragments.

The output of one pass through the shotgun sequencing method is a list of sequenced contigs. For the example of Figure 5.2, it would be the sequences of the three contigs shown. Of course, this does not complete the sequencing problem.

One can determine from the library of sequenced fragements alone neither the order nor the location of the contigs along S. For this one must return to the lab to find the locations of the contigs and to sequence the gaps.

Two mathematical problems must be solved to use the method effectively. First, what is an efficient and accurate algorithm for assembling fragments into contigs? Second, how well is S covered by the contigs for a given choice of N? The first question is a combinatorial and algorithmic problem which we shall not treat. We will assume it is solved and that we are able to link all overlapping fragments correctly. (Of course, this is a simplification ignoring sequencing errors and false linking of non-overlapping fragments that share a common end sequence by chance.)

We will consider the second issue, which is called the **coverage problem.** The coverage C attained by a shotgun sequencing of S is the proportion of S covered by contigs:

$$C \triangleq \frac{\text{total length of contigs}}{\text{total length of } S}. \tag{5.5}$$

Since shotgun sequencing is inherently random, C is a random variable. The expected value $E[C]$ is therefore a measure of how effectively shotgun sequencing covers a segment on average. The coverage problem is to construct a reasonable probabilistic model for the location and length of random fragments, and, given this model, to calculate $E[C]$ and show explicitly how it depends on N, L, and g. This is valuable information for shotgun sequencers because it tells them how to choose the number of fragments N needed to achieve the level of expected coverage they desire.

5.1.2 A coverage model.

Let S denote the segment of DNA we want to sequence and let g denote its length in basepairs. The first step of shotgun sequencing is shearing multiple copies of S into smaller fragments that can be sequenced individually. We shall assume that g is much larger than the typical fragment length. This is the most common situation and the one which shotgun sequencing is designed for.

We shall represent the segment S simply as the interval $[0, g]$, with the left endpoint 0 corresponding to the $5'$ end of S, the right endpoint g to its $3'$ end, and a point x in between 0 and g to the position along S which is x basepairs from the left ($5'$) end. A fragment of S is then some subinterval $[x, y]$ of $[0, g]$. Setting up the problem this way already entails an approximation. The position of the k^{th} base from the measured in basepairs is a whole number; real values which are not integers are not true positions and therefore S should really be represented by the sequence $\{1, 2, \ldots, g\}$. However, an arbitrary point in x in $[0, g]$ is at distance at most 1 from any point of $\{1, 2, \ldots, g\}$, a distance which is very small compared both to a typical fragment length and to g. Replacing $\{1, 2, \ldots, g\}$ by $[0, g]$ thus captures the essential picture, and $[0, g]$ is simpler to work with.

The next step in shotgun sequencing after producing a large pool of fragments of S is to randomly select N of them. Label these fragments by the integers from 1 to N. For each $1 \leq i \leq N$, let L_i denote the length of fragment i and X_i the position of its leftmost ($5'$) end. Both variables are in units of basepairs. The right endpoint of fragment i is then $Y_i = X_i + L_i$. (Actually, for integer-valued positions, $Y_i = X_i + L_i - 1$, but this difference is small relative to L_i and will be ignored.) Thus, fragment i is the subinterval $[X_i, Y_i]$ of $[0, g]$. Since the fragments are located randomly, X_i and L_i, $1 \leq i \leq N$, are random variables.

A probability model for shotgun sequencing will be a set of hypotheses about the joint distribution of $X_1, L_1, \ldots, X_N, L_N$. This joint distribution determines the distribution of C and hence determines $E[C]$, which is what we want to compute. The textbook model is based on the principles stated so far—g is much larger than fragment lengths, and the N fragments are obtained by random selection—plus two more: the fragmentation method is not biased toward any particular region of $[0, S]$, and the N fragments to be sequenced are selected from a population much larger than N. The three conditions that follow translate these principles into precise mathematical assumptions.

(i) For each fragment i, $1 \leq i \leq N$, X_i is uniformly distributed in the interval $(0, g)$; specifically, for any $0 \leq a < b \leq g$, $\mathbb{P}(a < X_i \leq b) = (b - a)/g$.

(ii) There is a maximum possible fragment length K for which $K \ll g$—this means K is much smaller than g, or equivalently, K/g is much less than one, and that $\mathbb{P}(L_i \leq K) = 1$ for all i.

(iii) The fragment lengths $L_1, \ldots, L_n$ all have the same probability distribution. All the random variables $X_1, \ldots, X_N, L_1, \ldots, X_N$ are independent.

Assumption (i) says, roughly, that each base in S is equally likely to be the left end of a fragment, in line with the idea that the fragmentation method favors no particular part of S. Assumption (ii) just states the hypothesis that g is much larger than fragment lengths. Assumption (iii) arises from combining two considerations. Since the N fragments are randomly selected, they are statistically independent, which means the pairs $(X_1, L_1), \ldots, (X_N, L_N)$ are independent. But if the fragmentation method is not systematically biased to any region of S, L_i should, in addition, be independent of X_i, for all i, and all fragment lengths should have the same distribution.

Some more approximations are hidden in this model. First, the fragments are really sampled without replacement, and so are not quite independent. But when the population is very large relative to sample size, the difference between sampling with replacement and without replacement is small, because then the probability of duplicates in a random sample with replacement is low. Second, L_i cannot really be independent of X_i because when X_i is close to g, $X_i + L_i$ must be less than g, which

constrains L_i. This is an 'edge' effect. We can imagine though the L_i is pretty much independent of X_i as long as $X_i < g-K$, where K is the maximum possible fragment length. Since X_i is uniformly distributed, the probability that $g - K \leq X_i < g$ is K/g, which is assumed to be small. So the probability of choosing a fragment in the troublesome zone can be neglected.

Observe that assumptions (i)—(iii) do not specify the probability distribution of the fragment lengths. Assumption (ii) is the only constraint. It turns out that the expected coverage in this model depends only on expected fragment length, so more details about fragment length distribution are unnecessary.

As first examples applying the model, we shall compute: a) the probability that fragment i covers a given position y in $(0, g)$; and b) the probability that a position y is in a gap between contigs. These results will help us compute expected coverage later.

Case I. Constant fragment length. First consider the unrealistic, but simpler case in which all fragments have a fixed length L, so that fragment length is not actually random.

Let y be a point of $[0, g]$, and assume $y \geq L$, so that an entire fragment can be fit between 0 and y. By definition, fragment i occupies the subinterval $[X_i, X_i + L]$ of $(0, g)$. It will cover y if and only if $X_i \leq y \leq X_i + L$, which happens if and only if $y - L \leq X_i \leq y$, as in Figure 5.3. But since X_i is uniformly distributed on $(0, g)$,

$$\mathbb{P}\left(\text{fragment } i \text{ covers } y\right) = \mathbb{P}\left(y - L \leq X_i \leq y\right) = \frac{y - (y - L)}{g} = \frac{L}{g}. \tag{5.6}$$

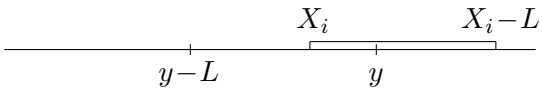

Figure 5.3. Covering y.

Position y is in a gap between contigs if none of the N fragments covers position y. From the previous equation, the probability that fragment i does not cover y is $1 - (L/g)$. Since the fragments locations are independent,

$$\mathbb{P}\left(\text{none of the } N \text{ fragments covers } y\right) = \left(1 - \frac{L}{g}\right)^N. \tag{5.7}$$

Case II. The general case. Now we let the fragment lengths be random. By assumption (iii), they all have the same distribution and therefore $E[L_i]$ is the same for all i. To show that the mean does not depend on i, we will denote it $E[L]$; thus, $E[L_i] = E[L]$ for all i.

To do this case it helps to review some probability theory. Let A be an event and define the random variable

$$\mathbf{1}_A = \begin{cases} 1, & \text{if } A \text{ occurs;} \\ 0, & \text{if not.} \end{cases}$$

This is a Bernoulli random variable and is called the indicator of A, because it indicates whether or not A has occurred. The expectation of a Bernoulli random variable is just the probability it equals 1. Thus

$$E[\mathbf{1}_A] = \mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A). \tag{5.8}$$

This is true for a conditional probability as well: if Z is a random variable, $\mathbb{P}(A|Z) = E[\mathbf{1}_A|Z]$. The purpose of these equalities is to use conditioning. Theorem 9 of Chapter 2 implies that if X and Z are two random variables, then

$$E[X] = E[E[X|Z]].$$

By applying this in (5.8),

$$\mathbb{P}(A) = E[\mathbf{1}_A] = E[E[\mathbf{1}_A|Z]] = E[\mathbb{P}(A|Z)]. \tag{5.9}$$

Now let us apply these formulae when A is the event that $[X_i, X_i+L_i]$ covers point y, and Z is replaced by L_i. Assumption (iii) says that L_i and X_i are independent, so, given L_i, X_i is still uniformly distributed on $(0, g)$. Thus, by repeating the derivation of (5.6), but with a fixed value of L_i replacing L,

$$\mathbb{P}\left(\text{fragment } i \text{ covers } y \middle| L_i\right) = \mathbb{P}\left(y - L_i \le X_i \le y \middle| L_i\right) = \frac{y - (y - L_i)}{g} = \frac{L_i}{g}. \tag{5.10}$$

We have assumed here that $y \ge K$, so that, since $L_i \le K$ no matter what, $y - L_i \ge 0$. Now apply (5.9):

$$\mathbb{P}\left(\text{fragment } i \text{ covers } y \middle| L_i\right) = E\left[\mathbb{P}\left(\text{fragment } i \text{ covers } y \middle| L_i\right)\right] = E\left[\frac{L_i}{g}\right] = \frac{E[L]}{g}. \tag{5.11}$$

Since the fragment locations are independent of another, the probability y is in a gap can be computed as before:

$$\mathbb{P}\left(\text{none of the } N \text{ fragments covers } y\right) = \left(1 - \frac{E[L]}{g}\right)^N. \tag{5.12}$$

The Clarke-Carbon formula for expected coverage.

We will show that under Assumptions (i)—(iii) above, the expected coverage is

$$E[C] \approx 1 - e^{-NE[L]/g}. \tag{5.13}$$

This is called the Clarke-Carbon formula. Several approximations are made in obtaining this result, which is why (5.13) is an approximate equality. The approximations will be explained in the derivation and they are all reasonable because K/g is assumed to be small.

The Clarke-Carbon formula is particularly nice because it expresses the coverage in terms of the single, intuitively meaningful number $a = NE[L]/g$, the ratio of the combined length of all the fragments to the length of S. It is standard to call a the coverage number. To do shotgun sequencing with $a\times$ coverage means to choose N to get coverage number a.

By definition, the expected value of C is $E[C] = \int_0^1 z f_C(z)\, dz$, where f_C is the density of C. But C is a rather complicated random variable whose density is not known, and so this expectation formula is of little direct use. We would like to compute $E[C]$ without first determining f_C. Fortunately there is a trick for doing this that works in a wide variety of circumstances. Expectation is a linear operation. As we showed in Chapter 2, equation (2.62), this implies

$$E\left[\int_a^b Z(r)\, dr\right] = \int_a^b E[Z(r)]\, dr.$$

If we can represent C as an integral of simpler random variables whose expectations we can compute, this formula will give us $E[C]$.

This can be done. For each location y in the interval $(0, g)$, define the Bernoulli random variable,

$$I(y) = \begin{cases} 1, & \text{if at least one of the } N \text{ fragments covers } y; \\ 0, & \text{otherwise.} \end{cases}$$

The function I indicates where the contigs are. Figure 5.4 shows the graph of I for a hypothetical outcome.

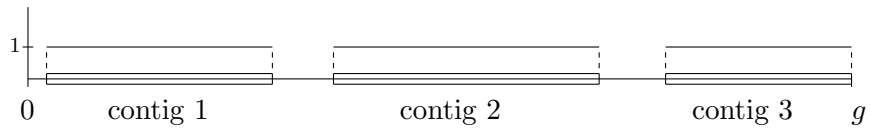

Figure 5.4: Graph of I.

From looking at Figure 5.4, it is clear that the total length of all the contigs is just the combined length of all those intervals where I has the value 1. Equivalently,

it is the area under the graph of I from $y = 0$ to $y = g$, which equals the integral $\int_0^g I(y)\,dy$. Thus,

$$C = \frac{1}{g}\int_0^g I(y)\,dy. \tag{5.14}$$

Moreover, since $I(y)$ is a Bernoulli random variable, $E[Y] = \mathbb{P}(Y=1) = 1-\mathbb{P}(Y=0)$. But $\{Y=0\}$ is just the event that no fragment covers y, and, in (5.12) we found its probability to be $(1-(E[L]/g))^N$. Thus

$$E[I(y)] = 1 - \left(1 - \frac{E[L]}{g}\right)^N.$$

Putting all this together,

$$\begin{aligned} E[C] &= \frac{1}{g}E\left[\int_0^g I(y)\,dy\right] = \frac{1}{g}\int_0^g E[I(y)]\,dy \\ &\approx \frac{1}{g}\int_0^g 1 - \left(1 - \frac{E[L]}{g}\right)^N dy = 1 - \left(1 - \frac{E[L]}{g}\right)^N. \end{aligned} \tag{5.15}$$

Finally, $E[L]/g$ is very small, so, by the approximation (5.4) discussed in the introduction to this Chapter, $\left(1 - \frac{E[L]}{g}\right)^N \approx e^{-NE[L]/g}$. Using this in in expression for $E[C]$ above gives the Clarke-Carbon formula.

Notice that an approximation sign also appeared between the second and third integrals in (5.15). There is not an equality here because the formula we are using for $E[I(y)]$ is only valid if $y \geq K$, but we have used it in the third integral as if it were true for all y. The error committed by doing so is at most K/g, which we are assuming is very small.

Expected number of contigs. Let W be the (random) number of contigs in a shotgun sequencing trial. It is also interesting to know the average number of contigs, $E[W]$ as a function of N, and it happens that this can also be computed easily:

$$E[W] \approx Ne^{-a} = Ne^{-NE[L]/g}. \tag{5.16}$$

A trick similar to the one used to compute expected coverage is helpful again. The probability mass function of W can probably not be found, so instead we represent W as a sum of Bernoulli random variables and use the linearity of expectations. For each fragment i, $1 \leq i \leq N$, define the random variable Z_i so that $Z_i = 1$ if fragment i is the rightmost fragment of a contig, and $Z_i = 0$ otherwise. Each contig has one rightmost fragment and so contains only one fragment for which $Z_i = 1$. Thus the number W of contigs is $\sum_1^N Z_i$, and by linearity of expectations,

$$E[W] = \sum_1^N E[Z_i] = \sum_1^N \mathbb{P}(Z_i=1). \tag{5.17}$$

To complete the calculation, it remains only to calculate $\mathbb{P}(Z_i = 1)$. The event $\{Z_i = 1\}$ occurs if and only if X_j falls outside of the interval $[X_i, X_i + L_i]$ occupied by fragment i, for every other fragment j. Because fragment i is now just an interval of length L_i somewhere in $[0, g]$, and because X_j is uniformly distributed and independent of X_i,

$$\mathbb{P}\left(X_j \text{ not in } [X_i, X_i + L] \middle| L_i\right) = 1 - \frac{L_i}{g},$$

and thus

$$\mathbb{P}\left(X_j \text{ not in } [X_i, X_i + L]\right) = E[1 - \frac{L_i}{g}] = 1 - \frac{E[L]}{g}.$$

The event that X_j is not in $[X_i, X_i + L]$ must occur for every of the other $N - 1$ fragments and they are all independent. Thus $\mathbb{P}\left(Z_i = 1\right) = (1 - L/g)^{N-1} \approx e^{-a}$. This result does not depend on i. Substituting back into equation (5.17) yields

$$E[W] = E\left[\sum_1^N Z_i\right] \approx \sum_1^N e^{-a} = Ne^{-a},$$

as claimed.

5.1.3 Exercises

Exercise 5.1.1. A DNA segment of length 2×10^5 bp is to be shotgun sequenced with a method producing fragments of length 5×10^2. How many fragments should be sequenced to obtained 99.5% expected coverage?

Exercise 5.1.2. In the derivation of the Clarke-Carbon formula we used the approximation, $(1 - L/g)^N \approx e^{-NL/g}$. If $g = 2 \times 10^5$, $L = 5 \times 10^2$, and $N = 2 \times 10^3$, find the error incurred by this approximation.

Exercise 5.1.3. A commercial shotgun sequencing firm finds that the cost to it of sequencing each fragment is \$20. Moreover, each base not covered by shotgun sequencing of a segment costs \$1 from customer dissatisfaction. Give an expression for the average total cost to the firm of shotgun sequencing applied to a segment of length g, using N fragments of length L. For given L and g, find the number of fragments N that minimizes this cost.

Exercise 5.1.4. As we noted in the derivation of the Clarke-Carbon formula, the formula (5.12) is not valid for $0 < y < K$, but we used it as if it were. Show that the error made by this approximation is no more than K/g. It is not necessary to compute $E[I(y)]$ exactly for $0 < y < L$. Just use the fact that $0 < E[I(y)] < 1$ (why?).

Exercise 5.1.5. Repeat the analysis of expected proportion of coverage and mean number of contigs using the following discrete model. (Ignore end effects.) Replace the interval $(0, g)$ by the sequence of integers $1, 2, \ldots, g$, which label the location of the bases in the original DNA segment. Assume

(i) The fragments all have a fixed integer size L.

(ii) The left endpoint X_i of fragment i is drawn uniformly from the set of integers $\{1, 2, \ldots, g\}$.

(iii) The locations $X_1, \ldots, X_N$ are independent.

Find $E[C]$ and the expected number of contigs for this model. Follow the reasoning used in the text but replace integrals by sums where appropriate. Ignore end effects in your calculation.

Exercise 5.1.6. Let U be a continuous random variable uniformly distributed on $(0, g)$. Let V be uniformly distributed on the integers $\{1, \ldots, g\}$. Show that for any number b, $\left|\mathbb{P}(U \leq b) - \mathbb{P}(V \leq b)\right| \leq 1/g$.

Exercise 5.1.7. In the case that $L_i = L$ is constant for all i, calculate $E[I(y)]$ exactly for $0 < y < L$, and then derive an exact formula for $E[C]$.

Exercise 5.1.8. Assume that $L_i = L$ is constant for all i. In this problem we will be interested in computing $E[I(y)I(z)]$ where z and y are points in $(0, g)$. Since $I(y)I(z) = 1$ if both y and z are covered by fragments and since it equals 0 otherwise, $I(y)I(z)$ is a Bernoulli random variable and

$$E\left[I(y)I(z)\right] = \mathbb{P}\left(\{I(y)=1\} \cap \{I(z)=1\}\right).$$

We can compute the probability on the right-hand side by computing the probability of the complementary event

$$(\{I(y)=1\} \cap \{I(z)=1\})^c = \{I(y)=0\} \cup \{I(z)=0\}.$$

It helps to recall the inclusion-exclusion formula, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

a) Assume that $L \leq y$ and $y + L \leq z$. Show that

$$E\left[I(y)I(z)\right] = 1 - 2(1 - L/g)^N + (1 - 2L/g)^N.$$

b) Use an exponential approximation in part a) to show that for $L \leq y$ and $y + L \leq z$, $E[I(y)I(z)] \approx E[I(y)]E[I(z)]$. Show that this may be interpreted as saying that $I(y)$ and $I(z)$ are approximately independent.

c) Find $E[I(y)I(z)]$ if $L \leq y < z < y + L$.

Exercise 5.1.9. Let $g = 10^6$, and assume that L is uniformly distributed on the interval $[10^3, 3 \times 10^3]$. How many fragments are required to achieve 90% expected coverage?

Exercise 5.1.10. Let $y < z < g$ and suppose that $z - y < K$. Suppose L_i is a continuous random variable with density f_L. (Note that $f_L(y) = 0$ for $y \geq K$. Show that the probability that fragment i covers both y and z is $(1/g) \int_{z-y}^{K} \ell f_L(\ell)\, d\ell - ((z - y)/g)\mathbb{P}(L \geq z - y)$.

5.2 Restriction Enzyme Digests; Discrete Models

The purpose of this section is to prepare the reader for Poisson models of restriction enzyme digests. We explain what restriction enzymes are and how they are used and do some preliminary modeling.

5.2.1 Restriction enzyme digests

Restriction enzymes, also called **restriction endonucleases**, are enzymes which cleave DNA molecules. They occur naturally and come in a variety of chemically distinct types. When mixed with DNA, a restriction enzyme will cut the DNA only at certain sites marked by a short sequence of nuceotides called a **recognition sequence**, specific to the enzyme. For example, the enzyme $ALuI$ recognizes the four letter sequence $AGCT$; if it encounters this sequence along one of the strands of the DNA double helix, it will cut the double helix between the G and the C nucleotides. Thus, $AluI$ will cut the sequence

$$AATGGCCTA\mathbf{AGCT}AGGGCTTC$$
$$TTACCGGAT\mathbf{TCGA}TCCCGAAG$$

into the pieces

$$AATGGCCTA\mathbf{AG} \quad \mathbf{CT}AGGGCTTC$$
$$TTACCGGAT\mathbf{TC} \quad \mathbf{GA}TCCCGAAG$$

Notice that the recognition sequence for $AluI$ is a reverse palindrome. Read backwards $AGCT$ becomes $TCGA$; replacing each base in $TCGA$ by its complementary base yields $AGCT$, which is the original sequence. In otherwords, the recognition site

$$\mathbf{AGCT}$$
$$\mathbf{TCGA}$$

will look exactly the sames after rotation by 180 degrees, so the same enzyme cleavage sites are specified by reading either strand. Reverse palindromic symmetry is a general (but not universal!) feature of restriction enzyme recognition sequences. If a restriction enzyme cuts a DNA strand between sites i and $i+1$ we shall call site i a **cut site** for that enzyme. Thus the cut site in the example is the eleventh site from the left.

Biologists have found a lot of clever ways to exploit restriction enzymes to obtain quantitative data in sequencing and in genetical studies of DNA. In general, they extract this data from a procedure called a **restriction enzyme digest.** Starting with a target segment of DNA, they clone it to produce many identical copies and then mix it with a restriction enzyme. The enzyme is allowed to act for awhile and

then is neutralized or washed out, leaving a solution of fragments of the original segment. If the enzyme is allowed to act for a sufficiently long time, it will cut the DNA at every restriction sequence that occurs along its length. This result is called a *complete digest* and breaks the DNA entirely into fragments extending between successive recognition sequences. Notice that the pool of fragments produced by a complete digest is quite different than that imagined in the shotgun sequencing method, because two fragments will be either identical or disjoint. If the restriction enzyme is not allowed to act for the full time, the result is a *partial digest* in which some recognition sequences are left uncut. Each fragment will extend between two recognition sequences but may contain undigested recognition sequences in the middle. Therefore a partial digest will, on average, produce longer fragments, and different fragments can overlap. The overlaps cannot be arbitrary but must extend between cuts at recognition sequences. The biologist may also take the fragments from one digest, whether complete or partial, and digest these with a second restriction enzyme having a different recognition sequence, thereby cutting the fragments into yet smaller pieces. This is a **double digest.**

The recognition sequences of various restriction enzymes may be found in textbooks on molecular biology or genomics. Here is a small selection to give some idea of the range of possibilities. The restriction enzyme $BamHI$ has recognition sequence $GGATCC$; it digests the double stranded segment

$$\begin{array}{l} 5'-GGATCC-3' \\ 3'-CCTAGG-5' \end{array} \quad \text{into the pieces} \quad \begin{array}{ll} 5'-G & GATCC-3' \\ 3'-CCTAG & G-5' \end{array}.$$

As you can see, it does not split the DNA clean through at one site, but leaves the strands on both sides with overhanging ends. This is actually typical of how restriction enzymes cut. Another restriction enzyme is $Hinfl$. Its recognition sequence $GANTC$, where N can be any one of the four nucleotide bases—in effect, $Hinfl$ has four recognition sequences. It also cuts double-stranded DNA so as to leave an overhang. As a last example, the restriction enzyme $HgiCII$ has the recognition sequence $GGWCC$ where W stands for either A or T. REBASE®, the Restriction Enzyme Data Base, available at `http://rebase.neb.com/rebase/rebase.html`, maintains up-to-date technical information on all the many restriction enzymes that have been identified by molecular biologists.

5.2.2 Mathematical preliminaries.

In this section we look at how to compute the probability that a given site in a DNA segment is the cut site for a given restriction enzyme. For convenience, denote this restriction enzyme by R, and let i be a given site, not too close to either end of the segment. To track whether i is a cut site or not, define the Bernoulli random variable

$$\xi_i := \begin{cases} 1, & \text{if site } i \text{ is a cut site of } R; \\ 0, & \text{otherwise;} \end{cases}$$

The probability i is a cut site for R depends on the probability model we use for the DNA sequence itself. Since the purpose of our analysis is rough estimates we will use the crudest and simplest model, the IID site model. This model was defined in Chapter 2, Example 2.8, but for convenience we repeat the definition here. Given a single strand of randomly drawn DNA, let X_i denote the base, either A, T, G, or C, at site i, where, as usual, sites are labeled starting from the $5'$-end. The IID site model assumes:

(i) $X_1, X_2, \ldots$ are mutually independent;

(ii) Each X_i has the same (identical) distribution, specified by the base probabilities p_A, p_T, p_G, p_C:

$$\mathbb{P}(X_i{=}A) = p_A, \quad \mathbb{P}(X_i{=}T) = p_T, \quad \mathbb{P}(X_i{=}G) = p_G, \quad \mathbb{P}(X_i{=}C) = p_C.$$

The IID site model with $p_A = p_T = p_G = p_C = 0.25$, is called the *IID site model with equal base probabilities.*

Computing the probability that i is a cut site is easy using this model because of independence. As an example, consider the restriction enzyme, *AluI*, and let

$$p \stackrel{\triangle}{=} \mathbb{P}\,(i \text{ is a cut site of } \mathit{AluI}) = \mathbb{P}(\xi_i{=}1).$$

(Of course we take $i \geq 2$, because $i = 1$ cannot be a cut site). Remember that i is a cut site if *AluI* can cleave the DNA sequence between sites i and $i+1$. Since the recognition sequence is $AGCT$, this can occur if and only if $X_{i-1} = A, X_i = G, X_{i+1} = C, X_{i+2} = T$. Assuming the IID sites model,

$$\begin{aligned} p &= \mathbb{P}\,(\mathit{AluI} \text{ cuts segment between sites } i \text{ and } i+1) \\ &= \mathbb{P}\,(X_{i-1}{=}A, X_i{=}G, X_{i+1} = C, X_{i+2} = T) \\ &= p_A p_G p_C p_T. \end{aligned} \tag{5.18}$$

Whatever the base probabilities p will be small and this will be the case for most any restriction enzyme.

More can be said. If i and j differ in position by at least 4 bases, the events that i and j are cut sites or not are independent, which is to say ξ_i and ξ_j are independent. This follows directly from the independence of sites, because whether or not i is a cut site depends on the bases at sites $i-1, i, i+1, i+2$, and whether or not j is depends on the bases at sites $j-1, j, j+1, j+2$, and these are disjoint if $|j-i| \geq 4$. If i and j are closer, ξ_i and ξ_j are not independent For example, if $\xi_i = 1$, then $X_{i-1}{=}A, X_i{=}G, X_{i+1} = C, X_{i+2} = T$ and it follows that $\xi_j = 0$, if $j \neq i$ and $i{-}3 \leq j \leq i{+}3$. Despite this, we can treat ξ_i and ξ_j as independent to the first order of approximation, even when $|i - j| < 4$, because p is small. Indeed, suppose that $\xi_1', \xi_2', \ldots$ are *independent* Bernoulli random variable with parameter p equal to the probability of a site being a cut site. If i and j are any two sites, the probability

that both $\xi'_i = 1$ and $\xi'_j = 1$ is p^2, which is much smaller than p. Thus, in a segment of moderate length, there will rarely be two sites within a distance 3 of each other for which both $\xi'_i = 1$ and $\xi'_j = 1$. Most possible sequences will have about the same probability that one would get treating $\xi_1, \xi_2, \ldots$ as if they were independent.

We will adopt this approximate viewpoint from now on and assume that whether a site is a cut site or not is approximately independent from site to site. We make this assumption for any restriction enzyme. This has two consequences for the model studied in the next section. First, the number of cut sites in the subsegment n bases long is (approximately) a binomial random variable with parameters n and p, where p is the probability that a site is a cut site. Second, the number of cut sites in disjoint portions of a DNA segment are independent of each other.

5.2.3 Problems

Exercise 5.2.1. Assume the IID site model with equal probabilities. Consider a restriction enzyme whose recognition sequence is ℓ bases long. Show that the probability that the enzyme cuts between site i and $i + 1$ is $(1/4)^\ell$.

Exercise 5.2.2 Assume the IID site model with equal probabilities. What is the probability that a site is a cut site for the restriction enzyme $HgiCII$ (see Section 5.2.1 for its recognition sequence)? Recompute p if instead $p_A = 0.2$, $p_T = 0.2$, $p_G = 0.3$, and $p_C = 0.3$.

Exercise 5.2.3. Show that the maximum of $p_A p_G p_T p_C$ over all parameters satisfying, $0 \leq p_A, p_G, p_T \geq 1$ and $p_A + p_G + p_C + p_T = 1$ is achieved at $p_A = p_G = p_C = p_T = 1/4$. (For example, use the Lagrange multiplier method.)

Exercise 5.2.4. Assume the cut probability is $p = .002$. (Assume the independence approximation.)

(a) Find an expression for the probability that in a segment of DNA there are 4 cuts in the first 1000 bases and at most two in the next 1000 bases.

(b) Find the probability that there are 6 cuts in the first 2000 bases and at least 3 are in the first 1000 bases.

(c) What is the expected number of cuts in the first 2000 bases?

Exercise 5.2.5. We showed that $\xi_1, \xi_2, \ldots$ are not really independent, and assuming them to be so is a modeling approximation that will introduce some error. However the independence assumption does not affect what the model predicts about the expected number of cuts in a sequence. Show that if we only assume $\xi_1, \xi_2, \ldots$ are all Bernoulli with the same parameter p, the expected number of cuts in a segment g base pairs long is gp, which is the same as when independence is assumed.

Exercise 5.2.6. Assume the i.i.d. site model for DNA sequence samples from an organism. What is the expected number of bases until the first occurrence of A in a randomly drawn sequence? What is the expected number of bases until the first occurrence of an A or a T (the pyrimidines)?

5.3 Poisson models and Restriction Enzyme Digests

5.3.1 The law of small numbers

The law of small numbers is a result about the binomial random variable with parameters n and p, when p is small and n is on the order of $1/p$. It says that for small p, large n, and moderate np, binomial distributions are well-approximated by Poisson distributions.

To see how the Poisson approximation arises, let us examine $\mathbb{P}(X=3)$ when X is a binomial random variable. According to the binomial probability formula

$$\mathbb{P}(X=3) = \frac{n!}{3!(n-3)!}p^3(1-p)^{n-3} = \frac{n(n-1)(n-2)}{3!}p^3(1-p)^{-3}(1-p)^n.$$

Define a new parameter $\lambda = np$. Since $p = \lambda/n$,

$$\begin{aligned}\mathbb{P}(X=3) &= \frac{n(n-1)(n-2)}{3!}\frac{\lambda^3}{n^3}(1-\frac{\lambda}{n})^{-3}\left(1-\frac{\lambda}{n}\right)^n \\ &= \left[\frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)}{\left(1-\frac{\lambda}{n}\right)^3}\right]\frac{\lambda^3}{3!}\left(1-\frac{\lambda}{n}\right)^n.\end{aligned}$$

Now assume that λ is of moderate size and p is small. Then n is large and $(1-\lambda/n)^n \approx e^{-\lambda}$, (see the introduction to this chapter). Also, since n is large $(1-1/n)(1-2/n) \approx 1$ and also $(1-\lambda/n) \approx 1$. Putting all this together,

$$\mathbb{P}(X=3) \approx \frac{\lambda^3}{3!}e^{-\lambda}, \qquad \text{where } \lambda = np. \tag{5.19}$$

The right-hand side is the probability that a Poisson random variable with parameter λ equals 3.

A similar analysis applies to $\mathbb{P}(X=k)$ for any positive integer k. For sufficiently large n, $\mathbb{P}(X=k) \approx (\lambda^k/k!)e^{-\lambda}$, where $\lambda = np$. In other words, a Poisson random variable with mean $\lambda = np$ is a good approximation of a binomial random variable with parameters n and p, when n is large and p is small enough that $\lambda = np$ is "moderate." Notice here that λ is the expected value of the Poisson random variable and np is the expected value of the binomial, so we are just approximating the binomial with a Poisson of equal mean.

There are several, precise quantitative formulations of the Poisson approximation. The first is a limit statement.

Theorem 1 *For each positive integer n, let X_n be a binomial random variable with parameters n and $p_n = \lambda/n$. Then for any non-negative integer k,*

$$\lim_{n\to\infty} P\left(X_n = k\right) = \lim_{n\to\infty}\binom{n}{k}(\lambda/n)^k\left(1-(\lambda/n)\right)^{n-k} = \frac{\lambda^k}{k!}e^{-\lambda}. \tag{5.20}$$

The second formulation establishes a bound on the error made by approximating a binomial with a Poisson random variable.

Theorem 2 *Let X be a binomial random variable with parameters n and p. Let Z be a Poisson random variable with $\lambda = np$. Then for any set A of non-negative integers,*

$$| P(X \in A) - P(Z \in A) | \leq p$$

Note that the bound in Theorem 2 shows that the error of the Poisson approximation can be bounded solely in terms of p, without regard to the size of n.

Example 5.2. A lottery is set up so that players have to guess a sequence of 6 digits, each between 0 and 9. If a million people play, what is the probability that there are two winners?

Assume that each player has a one in a million chance of guessing the right number, since there are one million 6 digit numbers. Assume also that each player guesses independently from all other players. The number of winners is then a binomial random variable with $n = 10^6$ and $p = 10^{-6}$. Hence the number of winners is approximately a Poisson random variable with $\lambda = 10^6 \times 10^{-6} = 1$. The probability of exactly two winners is then approximately

$$\frac{1}{2!}e^{-1} = 0.184$$

By the bound in Theorem 2, the error made in using the Poisson approximation in is less than 10^{-6}; this is pretty small—we did not even bother to calculate the answer to this order of accuracy! To 3 decimal places, the approximation coincides with the exact answer. ◇

The proof of Theorem 2 is hard and we will be omitted However Theorem 1 is not hard to demonstrate. We need only follow the same tricks we used in the approximation of $\mathbb{P}(X=3)$ above. For general k and $np_n = \lambda$,

$$\begin{aligned} \mathbb{P}(X=k) &= \frac{n!}{k!(n-k)!} p_n^k (1-p_n)^{n-k} \\ &= \left[\frac{n(n-1)\cdots(n-k+1)}{n^k (1-\lambda/n)^k} \right] \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n . \end{aligned}$$

The expression in the square brackets equals

$$\frac{1}{(1-\lambda/n)^k}(1-1/n)(1-2/n)\cdots(1-(k-1)/n)$$

and it tends to 1 as $n \to \infty$. On the other hand $\lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}$. Putting these two limits together proves Theorem 1. ◇

Some facts about the Poisson distribution were summarized in Chapter 2. We repeat them here, with proofs using moment generating functions; see Chapter 2 for the definition and for applications of the moment generating function.

Let X be a Poisson random variable with parameter λ. The moment generating function of X is

$$M(t) = E[e^{tX}] = \sum_{j=0}^{\infty} e^{tj} \frac{\lambda^j}{j!} e^{-\lambda} = e^{\lambda(e^t-1)} \tag{5.21}$$

Since

$$M'(t) = \lambda e^t e^{\lambda(e^t-1)} \quad \text{and} \quad M''(t) = \lambda \left(\lambda e^t + 1\right) e^{\lambda(e^t-1)},$$

the first two moments of X and the variance of X are

$$\begin{aligned} E[X] = M'(0) = \lambda, \qquad & E[X^2] = M''(0) = \lambda(\lambda+1) \\ \mathrm{Var}(X) \;\; = \;\; & E[X^2] - (E[X])^2 = \lambda. \end{aligned} \tag{5.22}$$

Suppose Z is a second Poisson random variable with parameter μ and assume Z is *independent* of X. Then, using formula (2.62) of Chapter 2, the moment generating function of $X + Z$ is

$$M_{X+Z}(t) = M_X(t) M_Z(t) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

But this last expression is the moment generating function of a Poisson random variable with parameter $\lambda + \mu$, so by Theorem 8 of Chapter 2, $X + Z$ is Poisson with parameter $\lambda + \mu$. We have proved a very nice property:

Theorem 3 *If X and Z are independent Poisson random variables with respective parameters λ and μ, then $X + Z$ is Poisson with parameter $\lambda + \mu$.*

5.3.2 Exercises

Exercise 5.3.1. Compare the Poisson approximation with the exact binomial probability for the following cases, and compare the error to the bound stated in the Theorem 2.

(a) $\mathbb{P}(X = 2)$, where X is binomial with $n = 12$ and $p = .15$.

(b) $\mathbb{P}(X = 7)$, where X is binomial with $n = 10$ and $p = .1$.

(c) $\mathbb{P}(X = 4)$, where X is binomial with $n = 12$ and $p = .04$.

Exercise 5.3.2. Assume that the probability that a site is a cut site for *AluI* us $p = 1/256$. Using an appropriate Poisson approximation, calculate the probability that a DNA segment 2000 bp long has 6 cut sites for *AluI*.

Exercise 5.3.3. Generalize the moment generating argument to show the following. Let $X_1, X_2, \ldots, X_n$ be independent Poisson random variables with respective parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$. Show that $X_1 + \cdots + X_n$ is Poisson with parameter $\lambda_1 + \cdots + \lambda_n$.

5.3.3 The Poisson process; definition

The aim of this section is to introduce the Poisson process, the most fundamental random process in applied probability. We will then use the Poisson process to model the location of restriction enzyme sites along a DNA segment and analyze some simple problems concerning statistics of cut site locations. In later sections we will use the Poisson process to discuss coverage problems for restriction enzyme digest libraries.

The Poisson process in one dimension is a model for points randomly placed on a line or ray. This is called a *point process*. We shall consider the case in which points are laid down on the interval $[0, \infty)$. We can think of $[0, \infty)$ as a time line; a point laid down at position t_1 in $[0, \infty)$ then represents an event that occurs at time t_1. For example, the event might be the arrival of a customer at a queue or of a job at a service center. Because the queueing interpretation is standard, in our general discussion we shall refer to the randomly placed points as *arrivals*.

To a given point process we can associate a counting process N_t that for each t counts the number of arrivals that occur in interval $[0, t]$. Then, given any two positions (times) s and t, with $0 \leq s < t$, $N_t - N_s$ is the number of arrivals that occur in $(s, t]$. Thus, the process N gives a complete description of the location of all the arrivals.

Figure 5 shows a typical sample path of a counting process. The horizontal axis is the time axis and arrivals are marked on this axis by the symbol $\times$. The figure shows the plot of N_t, which counts the number of arrivals. N_t is zero until the time of the first arrival, at which time it jumps to 1; it remains at 1 until the next arrival, at which time it jumps to 2. Thus, the process N_t is piecewise constant, and, assuming there are no simultaneous arrivals, increases by exactly one at the time of each new arrival.

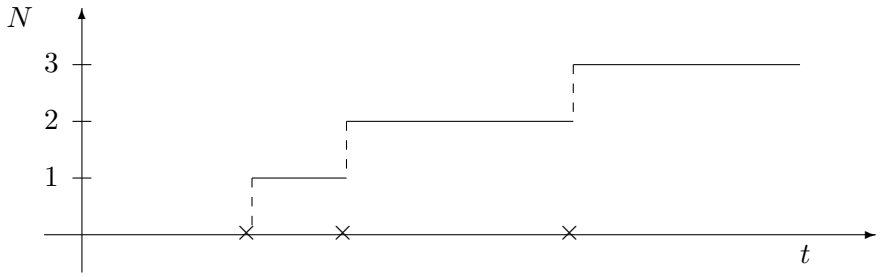

Figure 5.5: A point process and its counting process

A Poisson process is a counting process satisfying special conditions.

Definition. $N = \{N_t; t \geq 0\}$ is a Poisson process with *rate* ν if $N_0 = 0$ and

(P1) For every $0 \leq s < t$, $N_t - N_s$ is a Poisson *random variable* with parameter $\nu(t-s)$.

(P2) For each $0 \leq s < t$, $N_t - N_s$, the number of points falling in $(s,t]$, is independent of the values of N_u, $u \leq s$; in other words, the arrivals in $(s,t]$ are independent of those in $[0,s]$.

From assumption (P1) and the fact that the mean of a Poisson random variable with parameter λ is precisely λ,

$$E[N_t - N_s] = \nu(t-s),$$

if $t > s$. This expression is just the expected number of arrivals in the interval $(s,t]$. So the expected number of arrivals per unit time is

$$\frac{E[N_t - N_s]}{t-s} = \lambda,$$

which explains why λ is called the "rate" of the Poisson process.

The difference $N_t - N_s$ is called an increment of N because it measures the change in N over the interval $(s,t]$. Assumption (P1) implies that the probability distribution of the increment of N over $(s,t]$ depends only on its length $t-s$ and not on where the interval is located. This is called the *stationary increment property.* Assumption (P2) is called the *independent increment property,* because it states that increments of N over disjoint intervals are independent.

Let R be a restriction enzyme and let p denote the probability a site is a cut site. Consider a long strand of DNA (single-sided), let the parameter t measure distance in units of basepairs from its $5'$-end (at $t = 0$), and let N_t be the number of cut sites from site 0 to site t. N_t is a counting process. Assuming the IID site model for DNA, we observed in the last section that number of cut sites in a subsegment of length n is approximately binomial with paramters n and p. Since $N_t - N_s$ is the number of cut sites in the subsegment $(s,t]$ of length $t-s$, it follows that $N_t - N_s$ is approximately binomial with parameters $t-s$ and p. But p is small, and so if $t-s$ is of moderate or large size, then, making a second approximation, $N_t - N_s$ is approximately a Poisson random variable with parameter $\lambda = p(t-s)$. Thus $\{N_t\}$ satisfies condition (P1) approximately. We also observed in the last section that the number of cut sites in disjoint segments are independent random variables. This means that $N_t - N_s$ is independent of N_u whenever $0 < u \leq s < t$, since N_u counts the cut sites in $[0,u]$, which is disjoint from $(s,t]$. Therefore, N_t also *approximately satisfies* condition (P2) in the definition of a Poisson process. In short, $\{N_t\}$ is approximately a Poisson process. This motivates the following model.

Poisson process model for restriction enzymes. The process $\{N_t;\ t \geq 0\}$ counting restriction enzyme cut cites is a Poisson process with rate p, where p is the probability that any given site is a cut site.

Of course, this is an approximate model. Its great virtue is that it is very easy to work with.

Example 5.1. Suppose we are given a restriction enzyme 4 bases long, with $p = (1/4)^4$. What is the probability that there are exactly 5 cut sites in the first $(6)4^4$ base pairs of a DNA segment but that at most 1 cut site occurs in the first $(1.5)4^4$ base pairs?

We will answer this question treating N_t as a Poisson process with rate p. Notice carefully how we use properties (P1) and (P2) of a Poisson process. Let $s = (1.5)4^4$ and $t = (6)4^4$. The event whose probability we want to compute is the union of the two disjoint events:

$$\{N_s = 0, N_t - N_s = 5\} \cup \{N_s = 1, N_t - N_s = 4\}.$$

By the property $(P1)$, N_s is Poisson with parameter $\lambda_1 = sp = (1.5)4^4(1/4)^4 = 1.5$. Likewise, $N_t - N_s$ is Poisson with parameter $\lambda_2 = (t-s)p = ((6)4^4 - (1.5)4^4)(1/4)^4 = 4.5$. By property (P2), N_s and $N_t - N_s$ are independent. Thus

$$\begin{aligned} \mathbb{P}(N_s = 0, N_t - N_s = 5) &= \mathbb{P}(N_s = 0)\,\mathbb{P}(N_t - N_s = 5) \\ &= \frac{(1.5)^0}{0!}e^{-1.5}\frac{(4.5)^5}{5!}e^{-4.5} = \frac{(4.5)^5}{5!}e^{-6}. \end{aligned}$$

A similar calculation gives

$$\mathbb{P}(N_s = 1, N_t - N_s = 4) = \frac{(1.5)^1}{1!}e^{-1.5}\frac{(4.5)^4}{4!}e^{-4.5}.$$

Putting this together with some simplification yields

$$\begin{aligned} &\mathbb{P}(\{N_s = 0, N_t - N_s = 5\} \cup \{N_s = 1, N_t - N_s = 4\}) \\ &= \left(\frac{4.5}{5} + 1.5\right)\frac{(4.5)^4}{4!}e^{-6}. \qquad \diamond \end{aligned}$$

Poisson processes are standard models for counting the number of arrivals of customers or jobs to a queue. In fact, queueing theory is one of the most important domains of application of Poisson processes. The following examples further illustrate how to work with conditions (P1) and (P2).

Example 5.2. Jobs arrive at a server at a rate of 20/hour, according to a Poisson process. a) Find the probability that 15 jobs arrive in the first hour. b) Find the

probability that 5 jobs arrive in the first half hour and 10 jobs arrive in the second half hour. c) Given that no jobs arrive in the first 15 minutes, what is the probability that 4 arrive in the next 15 minutes?

Let N_t denote the jobs to arrive in the first t hours. The answer to the question a) is, using (P1),

$$\mathbb{P}(N_1 = 15) = \frac{(20)^{15}}{15!} e^{-20}.$$

The second question asks one to compute $\mathbb{P}(N_{.5} = 5, N_1 - N_{.5} = 10)$. By applying first the independence property (P2), and then the formula for Poisson probabilities,

$$\mathbb{P}(N_{.5} = 5, N_1 - N_{.5} = 10) = \mathbb{P}(N_{.5} = 5)\mathbb{P}(N_1 - N_{.5} = 10) = \frac{(10)^5}{5!} e^{-10} \frac{(10)^{10}}{10!} e^{-10}.$$

Question c) asks for $\mathbb{P}(N_{.5} - N_{.25} = 4 \;\big|\; N_{.25} = 0)$. However, by property (P2) of the Poisson process, the random variables $N_{.5} - N_{.25}$ and $N_{.25}$ are independent, and so

$$\mathbb{P}(N_{.5} - N_{.25} = 4 \;\big|\; N_{.25} = 0) = \mathbb{P}(N_{.5} - N_{.25} = 4) = \frac{5^4}{4!} e^{-5}. \qquad \diamond$$

Example 5.3. (Continuation of Example 5.1.) Assume the model of Example 5.1 with $p = 4^{-4}$.

a) What is the expected number of cut sites to arrive in the the segment of bases from the $5'$-end to base 3000?

b) What is the expected number of cut sites in $[0, 3000]$ given that there are 3 cut sites between base 0 and base 1000? hour, given that 14 jobs arrive in the first half hour?

Answering a) is just a matter of applying the formula $E[N(t)] = \lambda t$. Here $\lambda = p = 4^{-4}$ and $t = 3000$. Thus the expected number of cut sites is $3000/4^4$.

For b), we want to find $E[N_{3000} \mid N_{1000} = 3]$. To do this, first write $N_{3000} = N_{3000} - N_{1000} + N_{1000}$ and notice that, by the independent increment property, N_{1000} and $N_{3000} - N_{1000}$ are independent. Therefore

$$E[N_{3000} - N_{1000} \;\big|\; N_{1000} = 3] = E[N_{3000} - N_{1000}] = p(2000) = 2000/4^4.$$

and,

$$\begin{aligned} E[N_{3000}] &= E\left[N_{3000} - N_{1000} + N_{1000} \mid N_{1000} = 3\right] \\ &= E\Big[N_{3000} - N_{1000} \;\Big|\; N_{1000} = 3\Big] + 3 \\ &= \frac{2000}{4^4} + 3. \qquad \diamond \end{aligned}$$

5.3.4 Interarrival times (fragment lengths)

In this section we will study the time between arrivals of a Poisson process; if the Poisson process models location of cut sites, this interarrival time is the length of a segment between successive recognition sequences. The interrarival times will turn out to be exponential random variables. The reader not familiar with the exponential distribution should study the appropriate review material in section 2.2.4 of Chapter 2. We recall in particular that a random variable Z is exponential with parameter ν if and only if

$$\mathbb{P}(Z > s) = \int_s^\infty \nu e^{-\nu x}\, dx = e^{-\nu s}, \qquad \text{for all } s \geq 0. \tag{5.23}$$

N_t be a Poisson process with rate ν. In this section we will primarily think of t as a time parameter and suppose that N_t is counting arrivals by time t, but then we will reinterpret our results for the cut site location model.

Set $T_0 = 0$. Let T_1 be the time of the first arrival, T_2 the time of the second arrival, T_3 the time of the third arrival, and so on. Then, for each k, $T_{k+1} - T_k$ is the time that elapses between arrivals k and $k+1$ and is called an **interarrival time**. The time $T_1 = T_1 - T_0$, which is the time that elapses between 0 and the first arrival is thought of as the first interarrival time. We are interested in knowing the distributions and joint distributions of the arrival times and the interarrival times. We have first the following amazing result, which is so important, that we dignify it with the label of a theorem.

Theorem 4 *The interarrival times $T_1, T_2 - T_1, T_3 - T_2, \ldots$ of a Poisson process with rate ν are independent, identically distributed exponential random variables with parameter ν.*

Recalling that the expected value of an exponential with parameter ν is $1/\nu$, Theorem 4 implies the expected time between arrivals is

$$E[T_{k+1} - T_k] = \frac{1}{\nu}.$$

This makes perfect sense. The expected number of arrivals in any interval of length t is $E[N_t] = \nu t$; thus we should expect one arrival every $1/\nu$ units of time.

It is easy to see that T_1 is an exponential random variable, To say that $T_1 > t$ is precisely the same as saying $N_t = 0$. Thus,

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N_t = 0) = e^{-\nu t}.$$

Hence, by (5.23), T_1 is exponential with parameter ν.

We will omit the proof of the rest of the Theorem 4, except to give the following intuition. The event $T_i = s$ is an event concerning the Poisson process on $[0, s]$, so

it is independent of the process, $\{N_t - N_s;\, t > s\}$ of arrivals in $(s, t]$ for $t > s$. Now, given $T_1 = s$, $T_2 - T_1 > t$ only if $T_2 > s + t$, which happens only if there are no arrivals in $(s, s + t]$ or, equivalently, $N_{s+t} - N_s = 0$. But $N_{s+t} - N_s$ is a Poisson random variable with parameter $\lambda = \nu(t + s - s) = \nu t$ and is independent of what happens up to time s. Thus $\mathbb{P}(T_2 - T_1 > t | T_1 = s) = \mathbb{P}(N_{s+t} - N_s) = e^{-\nu t}$. Since this does not depend on s, $T_2 - T_1$ is independent of T_1 and $\mathbb{P}(T_2 - T_1 > t) = e^{-\nu t}$, proving $T_2 - T_1$ is again exponential with parameter ν. This argument is only formal because $\{T_1 = s\}$ is an event with probability zero.

Example 5.4. What is the expected time to the fifth arrival of a Poisson process

The fifth job arrives at time

$$T_5 = T_1 + (T_2 - T_1) + (T_3 - T_2) + (T_4 - T_3) + (T_5 - T_4).$$

Each interarrival time has expected value $1/\nu$ and hence $E[T_5] = 5/\nu$. ◇

Example 5.5. (Application to restriction enzyme digests) The recognition sequence of enzyme *BamHI* is *GGATCC*. Assuming the IID sites model with equal base probabilities to calculate the probability that a site is cut, state the Poisson process model for the cut site counting process. Let $L_1, L_2, \ldots$ be the lengths of the successive fragments produced by a complete digest of a DNA segment by *BamHI*. What can one say about $L_1, L_2, \ldots$ as random variables? What is the average length of a fragment?

Since the recognition sequence of *BamHI* has six letters and we are assuming the IID sites model with equal base probabilities, the probability p that a site is a cut site is $p = (1/4)^6 = 0.000244$. Thus, if N_t denotes the number of cut sites in the first t base pairs of a randomly drawn DNA seqment, the Poisson process model says $\{N_t\}$ is a Poisson process with rate $(1/4)^6$.

The "arrivals" of the process $\{N_t\}$ are the cut sites of *BamHI*, so the "time" T_i of arrival i is just the site of the i^{th} cut on the DNA segment, counting from the $5'$ end. A complete digest actually cuts the DNA at every valid cut site. Counting fragments from left to right, fragment i of the digest is therefore the interval of the DNA from site T_{i-1} to T_i, with length $L_i = T_i - T_{i-1}$. The fragment lengths are thus the interarrival times. According to Theorem 4, $L_1, L_2, \ldots$ are i.i.d. exponential random variables with parameter $\lambda = p = (1/4)^6$. Since the expected value of an exponential random variable with parameter λ is $1/\lambda$, the average fragment length is

$$E[L_i] = \frac{1}{p} = 4^6 = 4096 \quad \text{bp.}$$

This answer is of course intuitively obvious; if p is the probability that a site is cut, we should expect on average to see one cut per $1/p$ sites, which is another way of saying the average distance between cuts is $1/p$. ◇

5.3.5 Poisson thinning

Let N_t be a Possion process with rate ν. Think of N as counting arrivals to a queue, say for getting into a hot new club. Suppose that a bouncer stands at the end of the queue; he lets in each arrival with probability μ and turns them away with probability $1-\mu$. He does this independently for each arrival. Now let M_t count the arrivals that are allowed in. M is called a thinned Poisson process. The arrivals that are not let in are counted by $N(t)-M(t)$. This too is a thinned Poisson process.

If we use Poisson processes to model cut sites of a restriction enzyme, then thinned Poisson processes will model the actual cuts in a DNA segment made by a partial digest. Remember that in a partial digest, one lets the enzyme work on the DNA only for a short time so that the enzyme does not actually cut at every possible cut site, that is, at every location of a recognition sequence. If we assume that a cut takes place at a recognition sequence with probability μ, independently of what happens at other recognition sequences, the actual cuts are a thinned Poisson process.

The next result is an example of why Poisson processes are really nice; it tells us that a thinned Poisson process is also Poisson.

Theorem 5 *(Poisson thinning) Let $\{N_t\}$ be a Poisson with rate ν. Let $\{M_t\}$ be obtained by thinning $\{N_t\}$. If μ is the probability that an arrival is thinned, then $\{M_t\}$ is a Poisson process with rate $\mu\nu$. At the same time $N_t - M_t$ is a Poisson process with rate $(1-\mu)\nu$, and M_t, $t \geq 0$, and $N_t - M_t$, $t \geq 0$, are independent processes.*

Discussion: We explain intuitively why this theorem is true. To show that $\{M_t\}$ is a Poisson process with rate $\mu\nu$ we need to show that it satisfies properties (P1) and (P2), with ν replaced by $\mu\nu$.

Property (P2) is easy. $M_{t+h} - M_t$ is obtained by thinning the $N_t - N_s$ original arrivals in $(s,t]$. But by property (P2) for $\{N_t\}$, these arrivals are independent of the arrivals in $[0,s]$, and since each arrival is thinned independently of all others, it follows that $M_t - M_s$ is independent of all arrival and thinning events up to time s. This means that the increment $M_t - M_s$ is independent of M_u, $u \leq s$.

It remains to see why $M_t - M_s$ has the Poisson distribution. We can see why this should be true by imagining "binomial thinning." A binomial random variable with parameters (n,p) counts the number of heads in n independent tosses where p is the probability of heads. Now imagine that every time we toss the time, if it comes up heads, then with probability $1-\mu$ we turn over the coin to get tails, and with probability μ leave it heads. We do this independently for each toss. Thus, we have thinned the original number of heads. It is clear that the new probability for heads for each toss is $p\mu$ and that the results are independent between tosses. Thus the number of heads in n tosses after thinning is a binomial random variable

with parameters n and $p\mu$. Thus, a thinned binomial is a binomial. Thinking of Poisson random variables as law-of-small-number limits of binomials, it follows at a very intuitive level that a thinned Poisson is also Poisson. We shall not try to make this more precise, but we hope it provides the right intuition.

Finally, Theorem 5 contains the claim that M and $N - M$ are independent. This is surprising, since the two processes stem from the same original stream of arrivals counted by N and they add up to N. The proof is not hard, but it is a little technical and will not be discussed here either.

Example 5.6. Partial Digests. Suppose *BamHI* (see Example 5.5) is used to do a partial digest. Assume that in the partial digest the cut site of a recognition sequence is digested with probability 0.5. Let M_t be the number of sites that the partial digest cuts in the first t base pairs of a random DNA segments. State a model for $\{M_t\}$. Determine the probability distribution and expected value of a fragment length under this model.

Let $\{N_t\}$ denote the number of cuts sites in all the recognition sequences appearing in the first t base pairs. From Example 5.7, we can model this as a Poisson process with rate $(1/4)^6$. Since $\{M_t\}$ is a thinned version of $\{N_t\}$ with $\mu = 0.5$, it can be modelled as a Poisson process with rate $(0.5)(1/4)^6 = (1/2)^{13}$.

The fragment lengths of the partial digest are the interarrival times of $\{M_t\}$. Thus they are i.i.d. exponential random variables with parameter $\lambda = (1/2)^{13}$. The average fragment length is $1/\lambda = 2^1 3 = 8192$ base pairs. $\diamond$

5.3.6 Summing Poisson processes

Suppose that a stream of arrivals is the sum of two independent Poisson streams. That is, there is a Poisson process M_t, $t \geq 0$, with rate ν_1, counting one stream, and a second Poisson process Q_t, $t \geq 0$, with rate ν_2, counting a second stream. The total number of arrivals is the sum $N_t = M_t + Q_t$. The next result is yet a further example of how nice Poisson processes are.

Theorem 6 *If M and Q are independent Poisson processes with respective rates ν_1 and ν_2, then $N_t = M_t + Q_t$, $t \geq 0$, is a Poisson process with rate $\nu_1 + \nu_2$.*

We can actually prove this Theorem rigorously and completely and the proof is once again an opportunity to stress the definition of a Poisson process by properties (P1) and (P2).

First we verify that N has property (P1) for $\nu = \nu_1 + \nu_2$. Note that

$$N_t - N_s = [M_t - M_s] + [Q_t - Q_s] .$$

Since M is Poisson with rate ν_1, $M_t - M_s$ is a Poisson random variable with parameter $\nu_1(t-s)$. Similarly, $Q_t - Q_s$ is Poisson with parameter $\nu_2(t-s)$. By assumption

$M_t - M_s$ and $Q_t - Q_s$ are independent. By Theorem 3 for the sum of independent Poisson random variables, $N_t - N_s$ is Poisson with parameter $(\nu_1 + \nu_2)(t - s)$.

Next we verify (P2) for N. Fix $0 \leq s < t$. Then, since M is Poisson and independent of Q, $M_t - M_s$ is independent of M_u and of Q_u for $u \leq s$. Since N is the sum of M and Q, it follows that $M_t - M_s$ is independent of N_u for $u \leq s$. Similarly, $Q_t - Q_s$ is independent of N_u for $u \leq s$. This implies that $N_t - N_s = M_t - M_s + Q_t - Q_s$ is independent of N_u for $u \leq s$.

Thus we have shown that $\{N_t\}$ has stationary, independent, Poisson distributed increments and hence is a Poisson process.

Example 5.7. Double digests. This example continues the application of Poisson process properties to restriction enzyme digest, as begun in Examples 4.7 and 4.8. Suppose we digest a DNA segment in two steps. First we partially digest it with *AluI* and then completely digest the fragments with *BamHI*. In the partial digest, the probability that a cut site is digested is 1/10. Assume equal base probabilities in the computation of the probability a site is a cut site. Let N_t denote the number of sites cut by the double digest in the first t base pairs of the DNA segment. Determine the average fragment length.

The counting process model for the sites cut by a partial digest with *AluI* is Poisson with rate $(1/10)(1/4)^4$, since the recognition sequence for *AluI* is 4 bp long. The model for the sites cut by the complete digest by *BamHI* is Poisson with rate $(1/4)^6$. The process $\{N_t\}$ counting all cuts of the double digest is the sum of these two. The two cut site processes are approximately independent—we will justify this in a moment—so $\{N_t\}$ is (approximately) Poisson with rate $\nu = (1/10)(1/4)^4 + (1/4)^6 = 6.35 \times 10^{-4}$.

The fragments lengths of the double digest are the interarrival times of the process $\{N_t\}$, and they are exponentially distributed with parameter $\nu = 6.35 \times 10^{-4}$. Hence the average fragment length is $1/\nu = 2^{13}/3 \approx 1575$ base pairs.

Why is it that the cut site counting processes of the two digests are approximately independent and why not completely independent? They are not fully independent because cut sites for the two different recognition sequences cannot be arbitrarily close together; generally, a site cannot be a cut site of two different recognition sequences. However, imagine that we know where the cut sites are for the digest by *AluI*. Let us look at the second digest conditional on this knowledge only. Knowing the cut sites of the *AluI* means we know locations of the word *AGCT*. However, since the probability that a site is a cut site for *AluI* is small, these locations are relatively rare and scattered. The remainder of the DNA segment is unknown and approximately independent of this knowledge; since the second digest is a digest on this remaining portion it is essentially independent of the first digest. Admittedly, this argument is very rough and intuitive; the important point is that cut site probabilities are small for both of the two enzymes. ◇

5.3.7 Problems

Exercise 5.3.4. Consider the i.i.d. site model for bases along a DNA segment. Assume that the probability of seeing an A is .3, of seeing a T is .3, of seeing a G is .2, and of a C is .2. What should be the rate for a Poisson model of the location of cut sites of EcoRI, which has recognition sequence $GAATTC$? On average, how long are the fragments of a complete digest and what is their probability distribution?

Exercise 5.3.5. Assume the i.i.d. site model with equal probabilities to calculate the probability that a site is a cut site.

a) Using a complete digest by *EcoRI* (see the previous exercise) what is the probability that a fragment is shorter than 2000 bases given that it is longer than 1200 bases?

b) Suppose we do a partial digest with *EcoRI*, such that each recognition cut site is independently cut with probability .6. What is the distribution and mean length of the fragments in this case?

c) The recognition sequence of *Hinfl* is $GGWCC$, where W can be either A or T. State a model for the count of sites cut in a complete double digest using *EcoRI* and *Hinfl*. What is the distribution and mean fragment length in this case?

Exercise 5.3.6. The following problem, without the weak attempts at additional humor in its statement, is taken from a recent exam given by the Society of Actuaries. Its solution is an application of the rules for computing probabilities and expectations of increments of Poisson processes.

Here is a mathematical model in the exciting new field of dinosaurian predator dynamics. Field tests and evidence from recent cinema support the following model for Tyrannosaurus foraging needs.

(i) Tyrannosaurus uses calories uniformly at a rate of 10,000/day. If stored calories reach 0, the animal dies.

(ii) Tyrannosaurus feeds exclusively on scientists and each scientist provides exactly 10,000 calories.

(iii) Tyrannosaurus can store calories without limit until needed.

a) Suppose that a Tyrannosaurus catches scientists at a Poisson rate of 1 per day. Given that he has 10,000 calories stored at present, calculate the probability that he dies within the next 2.5 days.

b) Calculate the expected number of calories Tyrannosaurus eats in 2.5 days.

5.3.8 Residual and Current Life

Consider a Poisson process modeling arrival times. Fix a time t and consider the interval between the last arrival before t and the first arrival after t. We will be

interested in studying this random interval. If instead of arrivals the Poisson process counts cuts of a restriction enzyme digest, and if t represents a site, this is equivalent to studying the fragment of the digest containing site t.

To aid our study, we need a mathematical preliminary. Recall from Chapter 2 that Y has the gamma distribution with parameters λ and r if its density has the form

$$f(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{if } x > 0; \\ 0, & \text{otherwise,} \end{cases}$$

where $\Gamma(r) \stackrel{\triangle}{=} \int_0^\infty x^{r-1} e^{-x}\, dx$. The reason we introduce this random variable is the following result: if X and Y are independent exponential random variables each having parameter λ, then $X + Y$ has the gamma distribution with parameters λ and $r = 2$. It is easy to prove this using moment generating functions! From the table in section 2.3.5, the moment generating functions of X and Y are

$$M_X(t) = M_Y(t) = \frac{\lambda}{\lambda - t}, \qquad \text{if } t < \lambda.$$

Since X and Y are independent, we can use equation (2.67) from Chapter 2 to conclude that

$$M_{X+Y}(t) = M_X(t) M_Y(t) = \frac{\lambda^2}{(\lambda - t)^2}.$$

But, again refering to the table in 2.3.5, this is precisely the moment generating function of a gamma random variable with parameters λ and $r = 2$. By Theorem 8 of Chapter 2, since $X + Y$ has the moment generating function of such a gamma random variable, it must have the gamma distribution.

Now fix a Poisson process with rate ν and consider a time $t > 0$. The **residual lifetime**, R_t, is the time that elapses from t until the next arrival strictly later than t. The **current lifetime**, C_t, is the time since the last arrival before t if there is an arrival in $[0, t]$; if not, $C_t = t$. (If an arrival occurs at t, $C_t = 0$.)

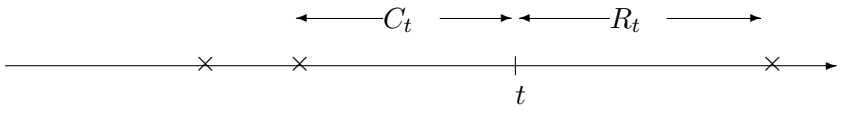

Figure 5.6: Current and Residual Lifetimes

The sum $C_t + R_t$ of residual and current lifetimes is the time between the last arrival before t and the first arrival after t. Ultimately, this sum is what we want to study. But first we look at R_t and C_t individually and as a pair.

The distributions of the residual and current lifetimes are easy to derive. The event that $R_t > s$ is the same as the event of no arrival in the interval $(t, t+s]$. Hence

$$\mathbb{P}(R_t > s) = \mathbb{P}\left(N(t+s) - N(t) = 0\right) = e^{-\nu s}. \tag{5.24}$$

This says that the residual lifetime is exponential with parameter ν, the same distribution as that of the interarrival times. It may seem strange that the residual lifetime and interarrival time are identically distributed, but this fact is really a reflection of the memoryless property of the exponential distribution (see Chapter 2, section 2.2.4). The distribution of the time to wait after t to see an arrival is independent of how much time has elapsed from the last arrival before t.

Consider now the current lifetime. It has what is called a truncated exponential distribution. Since arrivals are only counted starting from time 0, C_t cannot be larger than t; hence $\mathbb{P}(C_t > t) = 0$. But for $s \leq t$ the event that $C_t \geq s$ is just the event that there is no arrival in $(t-s, t]$, which has probability $\mathbb{P}(N_t - N_{t-s} = 0) = e^{-\nu s}$. To summarize,

$$\begin{cases} \mathbb{P}\left(C_t > s\right) = 0, & \text{if } s > t; \\ \mathbb{P}(C_t \geq s) = e^{-\nu s}, & \text{if } 0 \leq s \leq t. \end{cases} \tag{5.25}$$

Notice that

$$\mathbb{P}(C_t = t) = \mathbb{P}(C_t \geq t) - \mathbb{P}(C_t > t) = e^{-\nu t}.$$

We see that for $s < t$, this distribution looks precisely like the exponential distribution with parameter ν, but C_t can never be larger that t; hence the terminology "truncated exponential." The truncated exponential distribution is very close to a plain exponential distribution for large enough t, To understand what this means, consider the case in which $\nu = 1$ and $t = 8$. Then $t = 8$ is not even that large, but

$$\mathbb{P}(C_8 = 8) = e^{-8} \approx 0.0003 \approx .9997,$$

so with high probability C_8 falls in the interval $(0, 8)$, where its distribution follows the exponential exactly. *In what follows* we shall assume always that t is large enough that we model C_t approximately by a non-truncated exponential distribution.

Finally, it is easy to characterize the joint distribution of R_t and C_t. They are independent! The current lifetime C_t depends only on the arrivals that occur in $[0, t]$ and R_t depends only on the arrivals that occur in (t, ∞), and, for a Poisson process, the arrivals in these disjoint intervals are independent by the condition (P2).

Now we are able to characterize (approximately) the nature of $C_t + R_t$, the length of the interval between arrivals that contains t. We assume that t is large enough that C_t is approximately exponential with parameter ν. Since R_t is also exponential with parameter ν and C_t and R_t are independent, then, from the first calculation of this section $C_t + R_t$ *is approximately a gamma random variable* with parameters ν and $r = 2$. This is the main result we wanted to get to.

Example 5.8 Completely digest a DNA segment with *AluI*. What is the probability that the fragment containing site 3000 of the segment is at least 400 bp long?

Let N_t be the number of cuts sites for *AluI* in the first t bp of the DNA segment. Our model is that $\{N_t\}$ is Poisson with rate $p = (1/4)^4 = 1/256$. The length L of the fragment containg site 3000 is $L = C_t + R_t$ for $t = 3000$, and $C_t + R_t$ is approximately a gamma random variable with $\lambda = p = (1/256)$ and $r = 2$. Using the formula for the density of a gamma random variable and the fact that $\Gamma(2) = 1$,

$$\mathbb{P}(L \geq 400) \approx \int_{400}^{\infty} f_L(x)\,dx = \int_{400}^{\infty} p^2 x e{-}px\,dx.$$

Upon integration by parts,

$$\mathbb{P}(L \geq 400) \approx (p400 + 1)e^{-p400} = \left(\frac{400}{256} + 1\right) e^{-400/256} = 0.537. \quad \diamond$$

5.3.9 Application: Coverage probabilities for digest libraries

A digest library of a DNA segment is produced by digesting the segment with restriction enzymes and then sequencing the fragments of the digest. However, this method may miss some portions of the segment. Small fragments might get lost and other fragments might be too long to sequence. We model this by assuming that only fragments of length between two known levels $L < U$ can be sequenced. Digest fragments in this size range are called *sequencable*. The digest library will cover only that portion of the original DNA segment covered by sequencable fragments. The goal of this section coverage analysis. For given levels L and U and a given probability p that a site is cut by the digest, we want to calculate the expected proportion of the segment covered by sequencable digest fragments.

The analysis starts in a manner similar to the coverage analysis for shotgun sequencing. Assume the DNA segment to be digested is g base pairs long. For each position x along the segment, define

$$I_x \stackrel{\triangle}{=} \begin{cases} 1, & \text{if the length of the digest fragment containing } x \text{ is in } [L, U]; \\ 0, & \text{otherwise.} \end{cases}$$

I_x indicates whether x is in a sequencable fragment or not. The proportion of the segment covered by sequencable fragments is therefore $C = (1/g)\int_0^g I_x\,dx$, and the expected coverage is

$$E[C] = \frac{1}{g}\int_0^g E[I_x]\,dx = \frac{1}{g}\int_0^g \mathbb{P}(I_x = 1)\,dx. \tag{5.26}$$

To derive a formula for $E[C]$, we will compute $\mathbb{P}(I_x = 1)$, ignoring end effects as usual.

Let N_t, $0 \leq t \leq g$ be the process counting the sites along the DNA segment that are cut by the digest, and assume N is a Poisson process with rate ν. For a position

x along the segment, let R_x be the distance in bp from x to the next cut after x and let C_x be the distance between x and the last cut before x. The random variables R_x and C_x are the residual and current lifetimes at x, in the terminology of the previous section. Since $C_x + R_x$ is the length of the digest fragment containing x, x is in a sequencable fragment, that is, $I_x = 1$, if and only if

$$L \leq C_x + R_x \leq U.$$

But we know from the previous section that $C_x + R_x$ is approximately a gamma random variable with parameters $\lambda = \nu$ and $r = 2$, which has the probability density

$$f(y) = \nu^2 y e^{-\nu y}, \qquad y > 0.$$

So, using an integration by parts

$$\begin{aligned} \mathbb{P}(I_x = 1) &\approx \int_L^U \nu^2 y e^{-y\nu}\, dy \\ &= (\nu L + 1)e^{-\nu L} - (\nu + 1)e^{-\nu U}. \end{aligned}$$

This answer does not depend on x (ignoring end effects) and substitution into equation (5.26) gives the coverage formula:

$$E[C] \approx (\nu L + 1)e^{-\nu L} - (\nu U + 1)e^{-\nu U}. \tag{5.27}$$

Example 5.11. Optimizing a Partial Digest. Let N be the process counting the cuts sites of restriction sequences of a restriction enzyme along DNA, and suppose that N is modelled as a Poisson process with rate p, as usual. Suppose that the enzyme is used in a partial digest, with μ being the probability that a recognition sequence is cut. We have shown by Poisson thinning that the process M that counts the site actually cut is a Poisson process with rate $p\mu$. Given L and U, how should μ be chosen to obtain the best coverage?

We will choose μ to maximize the expected coverage. According to (5.27), for a digestion probability μ this is:

$$A(\mu) = (p\mu L + 1)e^{-p\mu L} - (p\mu U + 1)e^{-p\mu U}.$$

Now maximize $A(\mu)$ over the interval $0 \leq \mu \leq 1$ by using calculus; the details are left to an exercise. The result is that A is maximized at

$$\mu^* = \min\left\{ \frac{2}{(U-L)p} \log \frac{U}{L}, 1 \right\}. \qquad \diamond$$

5.3.10 Problems

Exercise 5.3.7. Show that $A(\mu)$ in the last calculation is maximized at μ^*.

Exercise 5.3.8. A digest library is built using two restriction enzymes. Enzyme I has a recognition sequence of length 6 and enzyme II has a recognition sequence of length 8. Only fragments whose length is between 1000 and 2000 bp's can be sequenced.

a) Find the expected proportion of coverage for a complete digest by enzyme I, a partial digest by enzyme I where the probability that a recognition sequence is cut is 0.7, and a complete double digest by enzymes 1 and 2.

b) Find the partial digest by enzyme I that has the best coverage.

Exercise 5.3.9. Let $\{N_t\}$ count the number of sites cut by a digest and assume it is Poisson with rate ν. Let x be a site along the DNA segment. What is the probability that x is more than K bp from the right end of the fragment in which it lies and less that J bp from the left end?