Chapter 5

Probabilistic analysis of sequencing methods

DNA carries hereditary information, via the genetic code, in the order in which nucleotide bases are arranged along its chains; this is the central fact about hereditary biology. Therefore, the problem of determining and interpreting DNA sequences is fundamental. Sequenced DNA from chromosomes or mitochondria provides basic data for studying genetics, evolutionary relationships between different organisms, and genetic diseases, and for searching for new genes and proteins. For this reason, DNA sequencing has become a big business. The technology has advanced to the point where it is possible to sequence entire genomes. Indeed, the International Human Genome Sequencing Consortium and Celera, Inc., under the direction of Craig Venter, announced working drafts of the entire human genome in the Spring of 2001. The human genome is approximately $3 \times 10^9$ bp long, so this was an ambitious project requiring massive amounts of lab work and data analysis.

In this chapter we discuss some mathematical problems that arise in a sequencing strategy called shotgun sequencing and also in sequencing the fragments of restriction enzyme digests of DNA. The solutions require building probabilistic models for experimental procedures and then analyzing the models mathematically, and they make a nice introduction to quantitative issues in sequencing. But, interesting as these applications are, we stress even more the modeling techniques and random processes behind them, which are used widely over a range of applications. For example, our study here of restriction enzyme digests is really an excuse to introduce the most important model of applied probability—the Poisson process. The Poisson process appears over and over again, in many fields, and is important in
random process theory as well. Using it to model fragments of a restriction enzyme digest is actually one of its lesser applications.

Everything that the reader needs to know about DNA and about restriction enzymes is explained in Chapter 1. The probabilistic models will employ continuous random variables, in particular uniform and exponential random variables. The reader should review Chapter 2, section 2.2.3, 2.2.4, and especially 2.2.5, which shows how continuous models arise as approximations to discrete random variables, as they will in this chapter. Expectations and moment generating functions, reviewed in Section 2.3, will also appear, but not the normal distribution.

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. Finally, Poisson random variables themselves arise as approximations of binomial random variables when the parameter $n$ is large, but $p$ is small, and Poisson processes are derived as approximations to discrete arrival processes. The reader should try to get a sense of when and why approximations are made. The purpose of the applications presented here is to give biologists numbers they can use as a guide to experimental design and interpretation of results, not to test a physical theory to arbitrary degrees of accuracy.

One mathematical fact from calculus will be used repeatedly and is the basis for the Poisson random variable approximations: for any real number $x$,

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

An informal statement of this fact is that

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

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? The following error bound for non-negative $x$ is a helpful guide:

$$\left| e^{-x} - \left( 1 - \frac{x}{n} \right)^n \right| \leq \frac{xe^{-x}}{2(1 - x/n)^2} \left( \frac{x}{n} \right) \quad \text{if } 0 \leq x < n. \quad (5.3)$$

At the same time, it is always true that $|e^{-x} - (1 - x/n)^n| < e^{-x}$, when $x \geq 0$, so for $x$ such that $e^{-x}$ is itself small, one should also look at the
relative error, \( |e^{-x} - (1 - x/n)^n|/(1 - x/n)^n \). For this we have a similar bound:

\[
\frac{|e^{-x} - (1 - x/n)^n|}{(1 - x/n)^n} \leq \frac{x^2}{2(1 - x/n)^2} \frac{e^{x^2/2n(1-x/n)^2}}{n}
\]

These bounds are not very accurate when \( x \) is of the order of \( n \). But they do show that when \( x \) is moderately sized and \( x/n \) is small, the approximation is good in both absolute and relative terms. For example, if we assume at least that \( x \leq 10 \) and \( n \geq 100 \). Then \( x/n \leq .1 \) and \( x^2/2n(1 - x/n)^2 \leq 1/1.62 \) and we can bound the absolute error by

\[
|e^{-x} - \left(1 - \frac{x}{n}\right)^n| \leq \frac{x}{1.6} \left(\frac{x}{n}\right),
\]

and the relative error by

\[
\frac{|e^{-x} - (1 - x/n)^n|}{(1 - x/n)^n} \leq \frac{x^2}{2(1 - x/n)^2} \frac{e^{x^2/2n(1-x/n)^2}}{n} \leq \frac{10 x}{1.62 n} e^{1/1.62} \leq \frac{12x}{n}.
\]

When \( n \to \infty \) in (5.3) while \( x \geq 0 \) is fixed, the right-hand side tends to zero, proving that \( (1 - x/n)^n \to e^{-x} \) as \( n \to \infty \), as claimed in (5.1). An easier proof of (5.1), which is valid for all \( x \), simply uses the representation \((1 - x/n)^n = e^{n \ln(1-x/n)}\) and the limit, \( \lim_{n \to \infty} n \ln(1 - x/n) = -x \). The reader is shown how to derive the bound in (5.3) in Exercise 5.2.

**Exercise 5.1.** Compute the bounds in (5.3) and (5.4) when \( n = 50 \), \( n = 100 \), and \( n = 400 \), for \( x = 1 \) and \( x = 5 \). Compute the exact absolute and relative errors in these cases and compare to the bounds.

**Exercise 5.2** a) Use the Mean Value Theorem of one-variable calculus to show that if \( z \geq y \geq 0 \), then \( |e^{-z} - e^{-y}| \leq e^{-y}|z - y| \).

b) \( T(y) \equiv -y \) is the first order Taylor polynomial centered at 0 for \( \ln y \). Use Taylor’s remainder theorem to show that for \( 0 < y < 1 \),

\[
|\ln(1 - y) + y| \leq \frac{y^2}{2(1 - y)^2}.
\]

c) Using \( (1 - x/n)^n = e^{n \ln(1-x/n)} \), and the fact that \(-\ln(1 - s) > s\) for \( 0 < s < 1 \), derive the bound in (5.3).

d) Note that if \( z \geq y \geq 0 \), then actually \( 0 < e^{-y} - e^{-z} < e^{-y} \), so in fact \( |e^{-y} - e^{-z}| < e^{-y} \). Use this to obtain from (5.3) the following bound, which is better over the whole range \( 0 \leq x < n \):

\[
|e^{-x} - \left(1 - \frac{x}{n}\right)^n| \leq e^{-x} \min\{1, \frac{1}{2(1 - x/n)^2} \frac{x^2}{n} \} \quad \text{if } 0 \leq x < n.
\]
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, we hasten to assure that shotgun sequencing does not actually 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. The segment length that a lab machine can handle changes with technological development, but, at least as of several years ago, was on the order of $5 \times 10^2$ to $2 \times 10^3$ base pairs. Yet biologists want to sequence DNA segments 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. It can be applied on different scales; we will explain it, not on the whole-genome scale, but for the problem of sequencing one moderately 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.

![Figure 5.1. Copies of $S$ differently fragmented.](image)
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.

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 $C_1$, $C_2$, and $C_3$ of $S$ that have been fully sequenced by connecting fragments with common ends.

![Figure 5.2: S with sequenced fragments, contigs, and gaps.](image)

The segments $C_1$, $C_2$ and $C_3$ are called **contigs**, from the word *contiguous*. They are contiguous stretches of linked, sequenced fragments which
Shotgun Sequencing 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 fragments 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 mainly 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:

$$C \triangleq \frac{\text{total length of contigs}}{\text{total length of } S}. \quad (5.5)$$

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.

In this section we present two standard models for random fragments in the shotgun sequencing method. The first model makes the simplifying assumption that fragment lengths are all of a fixed deterministic size; the second model is more realistic and allows random fragment lengths. In what follows, the letter $S$ will denote the long DNA segment being sequenced; for simplicity it is assumed that $S$ is single stranded. The letter $g$ will denote the length of $S$ in base pairs, $N$ will denote the number of sequenced fragments, and $L$ will be fragment length.
5.1.2 First coverage model.

We shall model the segment $S$, which has length $g$, 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. It is assumed that $g$ is a large number.

The shotgun sequencing method produces $N$ randomly located fragments. We label them with the integers from 1 to $N$. We represent the location of fragment $i$ as a subinterval $[X_i, Y_i]$ of $[0,g]$. Thus, $X_i$ denotes the position of the leftmost (5′ end) base and $Y_i$ denote the position of the rightmost base of fragment $i$. Since the fragments are located randomly, $X_i$ and $Y_i$, $1 \leq i \leq N$, are random variables. To create a probability model for the locations of the fragments only requires specifying the joint distribution of $X_1, Y_1, \ldots, X_N, Y_N$.

The first model makes the simplest possible assumptions:

Model I.

(i) Each fragment has identical length $L$, where $L \ll g$ (this means that $L$ is much less than $g$). Thus $Y_i = X_i + L$ for each $i$.

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

(iii) The fragment endpoint locations $X_1, X_2, \ldots, X_N$ are independent.

Remarks: The model makes a number of simplifying approximations. The first assumption, that the length $L$ is not random, is clearly unrealistic. But it makes the calculations easy, allowing us to derive a rough first estimate.

In the second assumption, the location $X_i$ of the left endpoint (the 5′ end) of a fragment is modelled as a continuous, uniform random variable. Assuming uniformity is saying that no one part of $S$ is more likely to be the site of a fragment end than another. It is a reasonable assumption if there is no reason to believe the fragmentation method is biased toward any region of $S$. However, assuming $X_i$ to be a continuous random variable looks plain wrong. The endpoint $X_i$ of a fragment occurs at a base sites along $S$, that is, at integer distances from the end of $S$. The “proper” model should be that $X_i$ is a discrete random variable uniformly distributed over the integers $\{1, 2, \ldots, g\}$, that is $P(X_i = k) = 1/g$ for each $1 \leq k \leq g$. Assumption (ii) amounts to approximating this discrete random variable by one that is continuous and uniformly distributed on $(0,g)$. Such approximations were
Shotgun Sequencing

discussed in Chapter 2 in reviewing continuous random variables. The reader
interested in the details should consult Section 2.3.2. But really, for this
chapter, one need only to accept the approximation as intuitively reasonable.
The main point is that if \( X \) is a continuous random variable uniformly
distributed on \((0, g)\), and if \( Y \) is a discrete random variable distributed on
\(\{1, \ldots, g\}\),
\[
\left| \Pr(X \leq b) - \Pr(Y \leq b) \right| \leq \frac{1}{g}, \quad \text{for every value } b.
\]
(5.6)
The left-hand-side of this inequality is the error made in using \( X \) to approx-
imate \( Y \) in a calculation of probabilities. When \( g \) is large—and typically it
is on the order of \(10^6\) or more—(5.6) shows the error made is vanishingly
small, and so modeling \( X_i \) by a continuous uniform random variable does
little harm. It is a useful approximation because it does makes the analysis
of the model easier.

There is a third, more subtle approximation implicit in assumptions (i)
and (ii). Suppose that \( X_i \) falls in the interval \((g - L, g)\). Then according
to (i) the right end of fragment \( i \) is at position \( X_i + L > g \), which is not
possible since \( g \) is the end of the segment. The same contradiction arises
with the initial fragment of the segment \( S \) if \( X_i \) happens to fall in \((0, L)\),
because then the fragment from position 0 to \( X_i \) has length \( X_i \), which is less
than \( L \). Strictly speaking, the assumption that \( X_i \) is uniformly distributed
on \((0, g)\) is inconsistent with the assumption that all fragments have equal
length, and we should modify assumption (i) to account for small fragments
at the end. We can do this simply by assuming the length of a fragment is
\( L \) unless \( X_i > g - L \), in which case the fragment is \((X_i, g)\) and has length
\( g - X_i \), and by ignoring initial fragments of length less than \( L \). But it is
not really necessary to worry about this. Assuming a uniform distribution,
the probability that \( X_i \) falls within a distance \( L \) of one of the endpoints is
just \(2L/g\). We are assuming that \( L \ll g \), so \( L/g \) is very small and these
short fragments occur with negligible probability and may be ignored in the
calculations. \( \diamond \)

As a first example applying the model, we shall compute the probability
that fragment \( i \) covers a given position \( y \) in \((0, g)\). We shall assume that
\( y \geq L \), 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 $(0, g)$,

$$
\Pr(\text{fragment } i \text{ covers } y) = \Pr(y - L \leq X_i \leq y) = \frac{y - (y - L)}{g} = \frac{L}{g}, \quad (5.7)
$$

![Figure 5.3. Covering y.](image)

The Clarke-Carbon formula for expected coverage.

Assuming Model I, we will derive the Clarke-Carbon formula for expected coverage:

$$
E[C] \approx 1 - e^{-NL/g}, \quad (5.9)
$$

Several approximations are made in obtaining the answer $1 - e^{-NL/g}$, which is why (5.9) is an approximate equality. The approximations will be explained in the derivation and they are all reasonable because $L/g$ 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 $a = NL/g$, 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 $a$.

We shall now derive the Clarke-Carbon formula. At first this might appear to be hard. 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 determining what this density is in order to compute the integral is potentially a difficult problem; it requires analyzing the probabilities of all different ways the fragments could be placed on $S$.
to get a certain coverage. Fortunately there is a trick for computing $E[C]$ without finding $f_C$. 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 of a shotgun sequencing.

![Figure 5.4: Graph of $I$.](image)

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 can be represented by the integral $\int_0^g I(y) \, dy$. It follows that,

$$C = \frac{1}{g} \int_0^g I(y) \, dy. \quad (5.10)$$

But this is great because, using linearity of expectations, as explained in Chapter 2, section 2.3.3,

$$E[C] = \frac{1}{g} E \left[ \int_0^g I(y) \, dy \right] = \frac{1}{g} \int_0^g E[I(y)] \, dy. \quad (5.11)$$

But $I(y)$ is a Bernoulli random variable, and from equation (5.8) we know that, independently of $y \geq L$,

$$E[I(y)] = P(I(y) = 1) = 1 - \left(1 - \frac{L}{g}\right)^N \quad (5.12)$$

Now this formula does not hold for $y$ when $0 \leq y < L$. But $[0, L]$ is a small interval compared to $[0, g]$, so we will fudge and use it for all $y$, because it will not affect the value of the integral in (5.11) much. So, plugging into (5.11),

$$E[C] \approx \frac{1}{g} \int_0^g 1 - \left(1 - \frac{L}{g}\right)^N \, dy = 1 - \left(1 - \frac{L}{g}\right)^N. \quad (5.13)$$
The approximation sign reminds us of our small fudge. In Exercise 5.6, the reader will see that the approximation makes an error only of the order of $L/g$, which we are assuming to be small.

The expression we wrote down for $E[C]$ in (5.13) is not yet the Clarke-Carbon formula stated in (5.9). The final step of the derivation requires one more approximation. Write $(1 - L/g)^N = (1 - (NL/Ng))^N$. We shall now use the approximation $(1 - x/N)^N \approx e^{-x}$, taking $x = NL/g$—see equation (5.2):

$$
\left(1 - \frac{NL}{N} \frac{1}{g}\right)^N \approx e^{-NL/g}.
$$

Substituting this into equation (5.13) gives the Clarke-Carbon formula as stated. Is the exponential approximation justified? We saw in Section 5.1, that $(1 - x/N)^N$ is nicely approximated by $e^{-x}$ when $x/n$ is small. But this is the case when $x = NL/g$ and $n = N$, for then $x/N = L/g$, and we are always assuming $L/g$ is 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 we can compute this easily for Model I. The answer turns out to be

$$
E[W] \approx Ne^{-a} = Ne^{-NL/g}.
$$

A trick similar to the one used to compute expected coverage works again in the calculation of $E[W]$. It is hard to derive the probability mass function of $W$, 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 $\sum_i^N Z_i$. So, by linearity of expectations,

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

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 $[X_i, X_i + L]$ occupied by fragment $i$, for every other fragment $j$. Because fragment $i$ is now just an interval of length $L$ somewhere in $[0, g]$, and because $X_j$ is uniformly distributed and independent of $X_i$,

$$
\mathbb{P}(X_j \text{ not in } [X_i, X_i + L]) = 1 - \frac{L}{g}.
$$
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
\[
\mathbb{P}(Z_i = 1) = (1 - L/g)^{N-1} \approx e^{-a}.
\]

This result does not depend on $i$. Substituting back into equation (5.15) yields
\[
E \left[ \sum_{i=1}^{N} Z_i \right] \approx \sum_{i=1}^{N} e^{-a} = Ne^{-a},
\]
as claimed in equation (5.14).

### 5.1.3 Second Coverage Model; random fragment lengths

The assumption that all fragments are of equal length is clearly unrealistic. Each fragment $i$, $1 \leq i \leq N$ will have its own length $L_i$, and, since fragments are chosen from a large pool of randomly broken DNA, these fragment lengths are actually random. In this section we state and analyze a simple model for random fragment lengths.

In proposing a model we want, as usual, to impose hypotheses that are as simple as possible, yet reasonable. We will continue to assume that the left ends $X_i$, $1 \leq i \leq N$, of the fragments satisfy the assumptions stated as (ii) and (iii) in the previous section. But instead of (i), we want to allow the fragment lengths $L_i$, $1 \leq i \leq N$ to be random variables and it is necessary to specify their joint distribution and their relation to $X_i$, $1 \leq i \leq N$. First, since the fragments are labeled in an arbitrary order, it is reasonable to suppose $L_1, \ldots, L_N$ are identically distributed. Since the $N$ fragments are in essence a random sample from a large fragment pool, it is also reasonable to assume the $L_1, \ldots, L_N$ are independent. And if the mechanical procedure shearing the original DNA segment into pieces works with the same intensity everywhere, $L_1, \ldots, L_N$ should also be independent of the fragment locations $X_1, \ldots, X_N$. Also, we expect all fragment lengths to be much smaller than the length $g$ of the segment $S$ being sequenced, and to enforce this, we will assume there is a fixed number $K \ll g$ such $L_i$ must be less than $K$. Finally, we shall assume that the $L_i$ are continuous random variables. To summarize, the new model replaces assumption (i) of the previous section by:

(i') The lengths $L_i$, $1 \leq i \leq N$ of the fragments are independent, identically distributed, continuous, positive random variables with a probability density function $f_L$. 

There is a number $K$, $K \ll g$, such that $P(L_i \leq K) = 1$. Equivalently $f_L(\ell) = 0$ for all $\ell \geq K$.

Finally, the fragment lengths $L_1, \ldots, L_N$ are independent of the fragment locations $X_1, \ldots, X_N$.

Recall that the joint density of independent random variables is the product of their densities; see Chapter 2, 2.2.6. Since by assumption (i'), $X_i$ and $L_i$ are independent and, by assumption (ii), $X_i$ is uniformly distributed on $(0, g)$, the joint density of $(X_i, L)$ is

$$g(x, \ell) = \begin{cases} \frac{1}{g} f_L(\ell), & \text{if } 0 < \ell < K \text{ and } 0 < x < g \\ 0, & \text{otherwise.} \end{cases} \quad (5.16)$$

We show that under assumptions (i)', (ii), and (iii), the following, modified Clarke-Carbon formula holds:

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

This is the same as the Clarke-Carbon formula of equation (5.9), except that $L$ has been replaced by $E[L]$.

The derivation of this formula follows that for the Clarke-Carbon formula step by step. Define the random variables $I(y)$ as before to be the Bernoulli random variable which indicates whether $y$ has been covered by one of the $N$ random fragments or not. We will calculate $E[I(y)]$, which is just the probability that $y$ is covered by at least one of the $N$ fragments.

Assume $y \geq K$. Then the event that $y$ is covered by fragment $i$ is $\{X_i \leq y \leq X_i + L_i\}$. According to the meaning of the joint density $g$ of $(X_i, L_i)$—see equation (2.43)—and using the fact that $g(x, \ell) = 0$ unless $0 < \ell < K$ and $0 < x < L$,

$$P(\{X_i \leq y \leq X_i + L_i\}) = \int_D \int g(x, \ell) \, dx \, d\ell \quad (5.17)$$

where $D$ is the region $\{(x, \ell); 0 < x < g, 0 < \ell < K, x \leq y \leq x + \ell\}$ in which $(X_i, L_i)$ must fall in order that fragment $i$ covers $y$. This region is shown in Figure 5.5.
By using the formula for $g$ in (5.16) and expressing the integral over $D$ as an iterated integral,

$$
P(\text{fragment } i \text{ covers } y) = \int_0^K \left[ \int_{y-K}^y f_L(\ell) \, dx \right] \, d\ell = \frac{1}{g} \int_0^K \ell f_L(\ell) \, d\ell = \frac{E[L]}{g}. \tag{5.18}
$$

Now the computation of $E[C]$ proceeds as before. The probability that fragment $i$ does not cover $y$ is $1 - E[L]/g$, and since the fragments are independent, the probability that none of the $N$ cover $y$ is $(1 - E[L]/g)^N \approx e^{-NE[L]/g}$. Thus $E[I(y)]$, the probability that at least one of the $N$ fragments covers $y$, is approximately $1 - e^{-NE[L]/g}$. Ignoring end effects, we have

$$
E[C] = \frac{1}{g} \int_0^g E[I(y)] \, dy \approx \frac{1}{g} \int_0^g \left( 1 - e^{-NE[L]/g} \right) \, dy = 1 - e^{-NE[L]/g}.
$$

### 5.1.4 Problems

**Exercise 5.3.** A DNA segment of length $2 \times 10^5$ bp is to be shotgun sequenced with a method producing fragments of length $5 \times 10^2$. How many fragments should be sequenced to obtained 99.5% expected coverage?

**Exercise 5.4.** 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 a bound on the error incurred by this approximation. Use the bound (5.3).
Exercise 5.5. 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 5.6. As we noted in the derivation of the Clarke-Carbon formula, the formula (5.6) is not valid for $0 < y < L$, but we used it as if it were. Show that the error made by this approximation is no more than $L/g$. 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 5.7. 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 5.8. Let $U$ be a continuous random variable uniformly distributed on $(0, g)$. Let $V$ be uniformly distributed on the integers $\{1, 2, \ldots, g\}$. Show that for any number $b$, $|\mathbb{P}(U \leq b) - \mathbb{P}(V \leq b)| \leq 1/g$.

Exercise 5.9. Assume Model I. Calculate $E[I(y)]$ exactly for $0 < y < L$, and then derive an exact formula for $E[C]$.

Exercise 5.10. Assume Model I. In this problem we will be interested in computing $E[II(y)I(z)]$ 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

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

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

$$\left(\{I(y) = 1\} \cap \{I(z) = 1\}\right)^c = \{I(y) = 0\} \cup \{I(z) = 0\}.$$
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
\[
E[I(y)I(z)] = 1 - 2(1 - L/g)^N + (1 - 2L/g)^N.
\]

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 5.11.** 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 5.12.** Assume model II. Let \( y < z < g \) and suppose that \( z - y < K \). 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 the models of restriction enzyme digests. We explain what restriction enzymes are and how they are used and we build a preliminary mathematical model. Our real goal for the rest of the chapter is to introduce the Poisson process and then use it as an approximate model for the location of restriction enzyme recognition sequences.

### 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 nucleotides 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

\[
AATGGCCTAAGCTAGGGCTTC \\
TTACCGGATTCGATCCCGAAG
\]
Restriction Enzyme Digests

into the pieces

\[ \text{AATGGCCTAAG AG CT} \]
\[ \text{AGGGCTTC TTACCGGAT TC GA} \]
\[ \text{TCCCGAAG} \]

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

\[ \text{AGCT} \]
\[ \text{TCGA} \]

will look exactly the same 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 \textbf{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 \textbf{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 \textbf{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 \textbf{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

\[ 5' - GGATCC - 3' \]
\[ 3' - CCTAGG - 5' \]

into the pieces

\[ 5' - G - GATCC - 3' \]
\[ 3' - CCTAG - G - 5' \].

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 Discrete models for cut site location. Mathematical framework.

The mathematical models for cut site location are applications of point processes and counting processes. In this section we introduce some basic notation defining such processes. We develop the notation directly as a model for cut site location, but the general set-up applies very widely, so it is worth understanding well.

Fix a restriction enzyme \( R \) and consider a DNA segment of length \( g \). We introduce the following random variables, indexed by the sites \( k, 1 \leq k \leq g \):

Define

\[ \xi_k \triangleq \begin{cases} 1, & \text{if site } k \text{ is a cut site of } R; \\ 0, & \text{otherwise}; \end{cases} \]

\[ M_k \triangleq \sum_{i=1}^{k} \xi_i. \]

Thus, the random variable \( \xi_k \) indicates whether \( k \) is a cut site or not, and \( M_k \) counts the number of cut sites up to site \( k \). We will be studying the sequences
of random variables, \( \{\xi_1, \xi_2, \ldots, \xi_g\} \) and \( \{M_1, M_2, \ldots\} \). We shall refer to these by the shorthand notations \( \{\xi_k\} \) and \( \{M_k\} \). Indexed sequences of random variables are called **stochastic processes** or **random processes**. In the context of restriction enzyme digest models, we will call \( \{\xi_k\} \) the cut site location process and \( \{M_k\} \) the cut site counting process.

As an example, consider the DNA sequence

\[ AAGGCCATAGCTTTTACCGTGTTGAGCTAAAGTTCTCCT, \]

and let us look at the cut site location and counting processes for the restriction enzyme AluI. The recognition sequence for AluI is AGCT, which it cuts between the G and the C. By examining this sequence, we see that sites 10 and 29 are the only cut sites. Therefore \( \xi_{10} = 1 \) and \( \xi_{29} = 1 \), and \( \xi_j = 0 \) for all other values of \( j \). As for \( \{M_k\} \), \( M_k = 0 \), if \( 1 \leq k \leq 9 \), since the first cut site is at site 10; \( M_k = 1 \) if \( 10 \leq k \leq 28 \), since the second cut does not occur until site 29; finally, \( M_k = 2 \), if \( 29 \leq k \leq 42 \). The information contained in the sequences \( \{\xi_1, \ldots, \xi_g\} \) and \( \{M_1, \ldots, M_g\} \) is represented graphically in the next figure. The \( x \)-axis represents the strand \( S \), the locations of the cut sites are marked by by the symbol ×, and the values of \( M \) are graphed with the symbol ◦.

![Figure 5.6](Image)

It is clear from the definitions and from the figure that the cut site location and counting processes are completely equivalent descriptions. In terms of the counting process \( \{M_k\} \), cut sites are located precisely at the indices at which \( M_k \) increases by 1. The counting process turns out to be the more useful than the location process for expressing the kinds of events we wish to study. For example, the number of cuts that occur in the segment from site \( k + 1 \) to site \( n \) is simply \( M_n - M_k \).

Counting processes are fundamental in queueing theory, which studies how networks of servers respond to a flow of arriving customers or jobs with service needs. In a queueing theory application, the \( x \)-axis in Figure 5.6,
representing a DNA segment, would be replaced by a time axis, and the
crosses would represent, for example, the arrival times of customers at a
queue. Then $M_k$ would represent the number of arrivals by time $k$. This
interpretation is very helpful for intuition and we shall return to it often.

5.2.3 Discrete models for cut site location: Bernoulli/binomial model

In this section we consider how to build a model for the cut site location
and counting processes. One way to get such a model is from a model of the
underlying DNA sequence. In chapter 2, we already introduced the IID sites
model for a random DNA sequence in the review of probability and random
variables. First, we will explore briefly the cut site location process that
results from the IID sites model. Then we will replace it by a much simpler
approximation—in fact, the approximate model is equivalent to independent
coin tossing! We will then show how to express this model in terms of the
cut site counting process.

The IID sites model for the bases along a strand of DNA, was defined
in Chapter 2, Example 2.9. We will repeat the definition here so it is not
necessary to go back to Chapter 2. 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 sites 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 $p_A, p_T, p_G, p_C$:

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

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.

We will show by example how the IID sites model determines the joint
distributions of the cut site location process for $AluI$. Let $i$ be a site. We
first calculate

$$p \triangleq P(i \text{ is a cut site of } AluI).$$

(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
Restriction Enzyme Digests

\[ X_{i-1} = A, X_i = G, X_{i+1} = C, X_{i+2} = T. \] Assuming the IID sites model,

\[ p = P(AluI \text{ cuts segment between sites } i \text{ and } i + 1) \]  (5.19)
\[ = P(X_{i-1} = A, X_i = G, X_{i+1} = C, X_{i+2} = T) \]  (5.20)
\[ = p_{APGCPT}. \]  (5.21)

We want to make a basic point here: whatever the base probabilities, \( p \) is small. For example, in the case of equal base probabilities—\( p_A = p_G = p_C = p_T = 0.25 \)—we have \( p = (0.25)^4 = 1/256 = 0.0039 \). But it can be shown that for any set of base probabilities \( p_A, p_G, p_T, p_C \),

\[ p_{APGCPT} \leq \left( \frac{1 \cdot 4}{4} \right)^4 = 0.0039, \]

so that 0.0039 is the highest possible value of \( p \) for \( AluI \), and indeed for any restriction enzymes with a recognition sequence 4 bases long. The reader is asked to show this in Exercise 5.15. Another simple and obvious point is that the cleavage probability \( p \) is independent of the site \( i \).

We can compute joint probabilities also. For example, it is clear that if \( i \) is a cut site, then \( i + 1 \) cannot be a cut site, so

\[ P(\xi_i = 1, \xi_{i+1} = 1) = 0, \quad P(\xi_i = 1, \xi_{i+1} = 0) = p. \]  (5.22)

Thus, \( \xi_i \) and \( \xi_{i+1} \) are dependent. On the other hand suppose that the distance between sites \( j \) and \( i \) is at least 4 base pairs. Then the sequences of sites \([i-1, i, i+1, i+2]\) and \([j-1, j, j+1, j+2]\) do not overlap. Since the bases at different sites are independent, it is clear that \( \xi_i \) and \( \xi_j \) are also independent. So, for example, \( P(\xi_i = 1, \xi_j = 1) = p^2 \).

What we have seen in these sample calculations generalizes to other restriction enzymes. Assuming the IID sites model, \( \xi_i \) and \( \xi_j \) will be independent except when the distance between \( i \) and \( j \) is less than the length of the recognition sequence. Since one is interested in studying the cut site processes over intervals much, much larger than the length of the recognition sequence, independence of sites is the dominant feature. Therefore, it is reasonable to propose a simplified, approximate model for \( \{\xi_i\} \) in which all dependence is suppressed.

**Model I, IID Locations:** \( \{\xi_k\} \) are i.i.d. Bernoulli random variables with some parameter \( p \).

We shall always compute \( p \) as we did in the example above, using the IID sites model. Notice that Model I is really just a coin tossing model in which the probability \( p \) of heads is typically very small.
Of course we know that for neighboring sites, Model I is just plain wrong. In reality, consecutive sites can never both be cut sites and so

\[ P(\xi_k = 1, \xi_{k+1} = 1) = 0. \]

But for model I, this “forbidden” event has positive probability \( P(\xi_k = 1, \xi_{k+1} = 1) = p^2. \)

So why should we even consider Model I as an approximation? The answer is that it works okay because \( p \) is very small. We know for recognition sequences of length 4 that \( p \leq 0.0039. \) In that case, \( p^2 \leq 1.53 \times 10^{-5}, \) which is much smaller yet. So in probability calculations using Model I, the assumption of independence between all sites will not introduce a large error. To put it another way, suppose we repeatedly and independently toss a coin with probability of heads equal to \( p. \) Then we will rarely see two heads in a row.

It will be very important of the development of Poisson process models in the following sections to restate Model I in terms of the counting process \( \{M_k\}. \) For this, it is convenient to include 0 as the first time index and set \( M_0 = 0. \) Then Model I is equivalent to two statements about the counting process.

\[ \text{M1} \text{ For each } 0 \leq j < k, \ M_k - M_j \text{ is binomial random variable with parameters } k - j \text{ and } p; \]

\[ \text{M2} \text{ For each } 0 \leq j < k, \ M_k - M_j \text{ is independent of } M_1, M_2, \ldots, M_j. \]

It is easy to see why these conditions are true if Model I holds. First,

\[ M_k - M_j = \sum_{i=1}^{k} \xi_i - \sum_{i=1}^{j} \xi_i = \sum_{i=j+1}^{k} \xi_i. \]

This is a sum of \( k-j \) independent Bernoulli random variables with parameter \( p, \) and hence is binomial with parameters \( k - j \) and \( p, \) as claimed in (M1). Second, \( M_1, \ldots, M_j \) are all defined in terms of \( \xi_1, \ldots, \xi_j, \) while, from the previous formula, \( M_k - M_j \) is defined in terms of \( \xi_{j+1}, \ldots, \xi_k. \) Since \( \xi_1, \ldots, \xi_j \) are all independent of \( \xi_{j+1}, \ldots, \xi_k, \) (M2) follows.

The converse statement, that assumptions (M1) and (M2) imply Model I, is left to Exercise 5.17.
Example 5.1 Consider Model 1, 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?

Let $s = 1.5(4^4)$ and $t = 6(4^4)$. The event in question is the union of the two disjoint events

$$\{M_s = 0, M_t - M_s = 5\} \cup \{M_s = 1, M_t - M_s = 4\}.$$ 

By the property (M1), $M_s$ is binomial with parameters $s$ and $p$ and $M_t - M_s$ is binomial with parameters $t - s$ and $p$. By property (M2), $M_s$ and $M_t - M_s$ are independent. Thus

$$\Pr(M_s = 0, M_t - M_s = 5) = (1 - p)^s \binom{t - s}{5} p^5 (1 - p)^{t - s - 5}.$$ 

A similar calculation gives

$$\Pr(M_s = 1, M_t - M_s = 4) = sp(1 - p)^s \binom{t - s}{4} p^4 (1 - p)^{t - s - 4}.$$ 

Putting this together with some simplification yields

$$\Pr(\{M_s = 0, M_t - M_s = 5\} \cup \{M_s = 1, M_t - M_s = 4\}) = \left(\frac{4.5(4^4)}{5}\right) + 1.5(4^4) \left(\frac{4.5(4^4)}{4}\right) (1/4)^{20}(1 - (1/4)^4)^6(4^4)^{-5}.$$ 

It’s not a very pretty answer, but we will learn how to get a nice approximation to such values using Poisson processes.

5.2.4 Problems

Exercise 5.13. Assume the IID site model with equal probabilities. Consider a restriction enzyme whose recognition sequence is $\ell$ bases long. Show that the probability that the enzyme cuts between site $i$ and $i + 1$ is $(1/4)^\ell$.

Exercise 5.14 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 5.15. 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
Exercise 5.16. Assume the cut probability is \( p = .002 \).

(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 5.17. Show that assumptions (M1) and (M2) above imply Model I.

Exercise 5.18. In the discussion of Model I we criticized the independence assumption. 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 5.19. In this problem we will work with the recognition sequence \( \text{AGCT} \) of the restriction enzyme \( \text{AluI} \). In our discussion criticizing the independence assumption in Model I, we observed the Model I assigns a positive probability to the event \( \xi_1 = 1, \xi_{i+1} = 1 \), whereas in reality this cannot happen. Then we argued that the probability assigned by Model I to this event is very small. In this problem, we pursue this issue further

a) Show that for \( \text{AluI} \) it is not possible to have more than one 1 among the four random variables \( \xi_i, \xi_{i+1}, \xi_{i+2}, \xi_{i+3} \). Find the probability under Model I that this "forbidden" event happens. (Note: the forbidden event is the same as \( \sum_{j=i}^{i+3} \xi_j \geq 2 \)).

b) Assuming model I, calculate the probability that \( \xi_i = 0, \xi_{i+1} = 0 \). (To avoid trivialities, assume \( i \geq 2 \)).

c) Assuming the IID sites model calculate the probability that \( \xi_i = 0, \xi_{i+1} = 0 \), and compare the answer to that in b).

Exercise 5.20. 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 an \( A \) in a randomly drawn sequence? What is the expected number of bases until the first occurrence of an \( A \) or a \( \text{T} \) (the pyrimidines)?
5.3 Poisson random variables and Poisson processes

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 \( p \) in the domain where \( p \) 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. This approximation considerably simplifies working with the counting process of a stream of Bernoulli arrivals.

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

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

Define a new parameter \( \lambda = np \). We will assume that \( \lambda \) is of moderate size and that \( p \) is small, so that \( n \) is large. Since \( p = \lambda/n \), we can rewrite \( P(X = 3) \) in terms of \( \lambda \) and \( n \):

\[
P(X = 3) = \frac{n(n-1)(n-2)}{3!} \frac{\lambda^3}{n^3} (1 - \frac{\lambda}{n})^3 \left( 1 - \frac{\lambda}{n} \right)^n
\]

\[
= \left[ \frac{(1 - \frac{1}{n}) (1 - \frac{2}{n})}{(1 - \frac{\lambda}{n})^3} \right] \frac{\lambda^3}{3!} \left( 1 - \frac{\lambda}{n} \right)^n.
\]

Since \( n \) is large and \( \lambda \) is of moderate size our favorite approximation applies (see the introduction to this chapter): \((1 - \lambda/n)^n \approx e^{-\lambda}\). Also, since \( n \) is large \((1 - 1/n)(1 - 2/n) \approx 1\) and also \((1 - \lambda/n) \approx 1\). Putting all this together,

\[
P(X = 3) \approx \frac{\lambda^3}{3!} e^{-\lambda}, \quad \text{where } \lambda = np. \tag{5.23}
\]

The right-hand side is the probability that a Poisson random variable with parameter \( \lambda \) equals 3.

A similar analysis applies to \( P(X = k) \) for any positive integer \( k \). For sufficiently large \( n \), \( P(X = k) \approx (\lambda^k/k!) e^{-\lambda} \), 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 \( \lambda \) 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 \( k \),

\[
\lim_{n \to \infty} P(X_n = k) = \lim_{n \to \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{5.24}
\]

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 example 2 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.

Theorem 2 is tricky and we will not attempt to show why it is true. However Theorem 1 is not hard to demonstrate. We need only follow the same tricks we used in the approximation of \( P(X = 3) \) above. For general \( k \) and \( np_n = \lambda \),

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

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 the limit in 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 \( \lambda \). The moment generating function of \( X \) is

\[
M(t) = E[e^{tX}] = \sum_{j=0}^{\infty} e^{tj} \frac{\lambda^j}{j!} e^{-\lambda} = e^{\lambda(e^t-1)} \quad (5.25)
\]

Since

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

the first two moments of \( X \) and the variance of \( X \) are

\[
E[X] = M'(0) = \lambda, \quad E[X^2] = M''(0) = \lambda(\lambda + 1) \quad (5.26)
\]

\[
\text{Var}(X) = E[X^2] - (E[X])^2 = \lambda.
\]

Suppose \( Z \) is a second Poisson random variable with parameter \( \mu \) 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 $\lambda$ and $\mu$, then $X + Z$ is Poisson with parameter $\lambda + \mu$.

### 5.3.2 Problems

**Exercise 5.21.** 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) $P(X=2)$, where $X$ is binomial with $n = 12$ and $p = 0.15$.

(b) $P(X=7)$, where $X$ is binomial with $n = 10$ and $p = 0.1$.

(c) $P(X=4)$, where $X$ is binomial with $n = 12$ and $p = 0.04$.

**Exercise 5.22.** Assume that the probability that a site is a cut site for AluI is $p = 1/256$. Using an appropriate Poisson approximation, calculate the probability that a DNA segment 2000 bp long has 6 cut sites for AluI.

**Exercise 5.23.** 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 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.

The relation between the point process and its counting process is precisely the same as that between the cut site location process and the cut site counting process defined in Section 5.2.2. The only difference is that now arrivals can occur at any time, not just at integer times, and the parameter $t$ in $N_t$ ranges continually through the nonnegative numbers, rather than being restricted to integer values. The next figure illustrates the relationship between the point process and it counting process. It is just a slight modification of Figure 5.6. This time the crosses mark arrivals, and the graph is that of the process $\{N_t\}$.

![Figure 5.7: A point process and its counting process](image)

A Poisson process is a counting process satisfying special conditions. As you read the definition, bear in mind the conditions $(M1)$ and $(M2)$ for the counting process $\{M_k\}$ of Model I.

**Definition.** $N = \{N_t; t \geq 0\}$ is a Poisson process with rate $\nu$ 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 $\lambda$ is precisely $\lambda$,

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

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

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

which explains why $\nu$ 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) is called the stationary increment property. In words, (P1) says 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. Assumption (P2) is called the independent increment property, because it states that increments of $N$ over disjoint intervals are independent.

To see what motivates the definition of a Poisson process think back to the conditions (M1) and (M2) for the counting process $\{M_k\}$ of the cut site location model. Recall that there we defined $M_t = \sum_1^t \xi_k$ for integer values of $t$, where $\xi_1, \xi_2, \ldots$ were i.i.d. Bernoulli random variables with parameter $p$. The condition (M1) was: for all integers $0 \leq s < t$, $M_t - M_s$ is binomial with parameters $(t - s)$ and $p$. But we know from the previous section that for small $p$ and moderate $p(t - s)$, this binomial random variable can be approximated by a Poisson random variable with parameter $\lambda = p(t - s)$. Assumption (P1) (with $\nu = p$) can be viewed as the consequence of applying the Poisson approximation to the increments of $\{M_k\}$. On the other hand, assumption (P2) is a direct generalization of (M2). For the counting process $M_t$, the arrivals $\xi_1, \xi_2, \ldots$ are independent and so the total number of arrivals between times $s$ and $t$ are independent of arrivals outside this interval. (P2) expresses exactly the same assumption.

Conditions (M1) and (M2) were stated in the particular context of a cut site location model. But there is no need to insist on this interpretation. Stated generally, we have shown that a Poisson process arises naturally from the Poisson approximation to the counting process $M_t = \sum_1^t \xi_i$ of an i.i.d. Bernoulli process $\{\xi_i\}$.

Since the cut site probability $p$ for a restriction enzyme is small, the Poisson approximation applies and allows us to replace Model I by a Poisson
process model. We state this now. We suppose we are given a restriction enzyme, and we let \( p \) denote the probability that it cuts at a site. Let the parameter \( t \) denote the distance along a DNA segment from the 5’-end; we let \( t \) vary continuously over the non-negative real numbers, as if it were a time variable, and we measure \( t \) in units of base pairs. Let \( N_t \) denote the number of recognition sequence cut sites in \([0, t]\); we use \( N_t \) instead of the previous notation \( M_t \) to emphasize we are using the Poisson model. The new model replacing (M1) and (M2) is:

**Model II.** \( \{N_t; \ t \geq 0\} \) is a Poisson process with rate \( p \).

To show how assumptions (P1) and (P2) work and the ease with which calculations can be made with them, we shall redo Example 5.1.

**Example 5.3.** (Example 5.1, revisited) Suppose we are given restriction enzyme with \( p = (1/4)^4 \). What is the probability that there are exactly 5 cut sites in the first \((6)^4\) base pairs of a DNA segment but that at most 1 cut site occurs in the first \((1.5)^4\) base pairs?

Let \( N_s \) denote the number of cut sites in \([0, s]\) and assume Model II. Let \( s = (1.5)^4 \) and \( t = (6)^4 \). 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 \( \lambda_1 = sp = (1.5)^4(1/4)^4 = 1.5 \). Likewise, \( N_t - N_s \) is Poisson with parameter \( \lambda_2 = (t - s)p = ((6)^4 - (1.5)^4)(1/4)^4 = 4.5 \). By property \((P2)\), \( N_s \) and \( N_t - N_s \) are independent. Thus

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

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

\[
\begin{align*}
\mathbb{P}(\{N_s = 0, N_t - N_s = 5\} \cup \{N_s = 1, N_t - N_s = 4\}) &= \left(\frac{4.5}{5} + 1.5\right) \frac{(4.5)^4}{4!} e^{-6}.
\end{align*}
\]
This is a prettier answer than the one derived in Example 5.1 and it is much easier to compute numerically.

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. To further illustrate how to work with conditions (P1) and (P2), we solve some problems framed in the language of arrivals.

Example 5.4. 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),

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

The second question asks one to compute $P(N_5 = 5, N_1 - N_5 = 10)$. By applying first the independence property (P2), and then the formula for Poisson probabilities,

$$
P(N_5 = 5, N_1 - N_5 = 10) = P(N_5 = 5)P(N_1 - N_5 = 10) = \frac{(10)^5}{5!} e^{-10} \frac{(10)^{10}}{10!} e^{-10}.
$$

Question c) asks for $P(N_5 - N_{25} = 4 \mid N_{25} = 0)$. However, by property (P2) of the Poisson process, the random variables $N_5 - N_{25}$ and $N_{25}$ are independent, and so

$$
P(N_5 - N_{25} = 4 \mid N_{25} = 0) = P(N_5 - N_{25} = 4) = \frac{5^4}{4!} e^{-5}.
$$

Example 5.5. (Continuation) For the set-up of the previous example:

a) What is the expected number of jobs to arrive in the first half hour?  
b) What is the expected number of jobs to arrive in the 30 minute interval from minute 10 to minute 40?  
c) What is the expected number of jobs to arrive in the first hour, given that 14 jobs arrive in the first half hour?
Answering a) is just a matter of applying the formula \( E[N(t)] = \lambda t \). In problem a), \( \lambda = 20 \) and \( t = 1/2 \). Thus the expected number of jobs to arrive in the first half hour is \( E[N_{1/2}] = 20/2 = 10 \) jobs.

The answer to b) is also 10. This is immediate from the stationarity property (i). The probability distribution of the number of jobs to arrive in any 1/2 hour interval is just the same as for the number to arrive in the first 1/2 hour.

In c), we want to find \( E[N_1 \mid N_{5} = 14] \). To do this first write \( N_1 = N_1 - N_{5} + N_{5} \) and notice that, by the independent increment property, \( N_{5} \) and \( N_1 - N_{5} \) are independent; in other words, the number of jobs that arrive in the first half hour has no effect on the expected number that arrive in the second half hour. Thus

\[
E[N_1 - N_{5} \mid N_{5} = 14] = E[N_1 - N_{5}] = 10.
\]

Therefore,

\[
E[N_1] = E[N_1 - N_{5} + N_{5} \mid N_{5} = 14] = E[N_1 - N_{5}] + E[N_{5} \mid N_{5} = 14] = 10 + 14 = 24.
\]

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 interarrival 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 \( \nu \) if and only if

\[
P(Z > s) = \int_s^{\infty} \nu e^{-\nu x} \, dx = e^{-\nu s}, \quad \text{for all } s \geq 0. \tag{5.27}
\]

The fact that the inequality in the event \( Z > s \) is strict is not important; if \( Z \) is exponential it is a continuous random variable and hence for any \( s \), \( P(Z = s) = 0 \); hence \( P(Z \geq s) = P(Z = s) + P(Z > s) = e^{-\nu s} \), as well.

\( N_t \) be a Poisson process with rate \( \nu \). 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 also considered to be 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 $\nu$ are independent, identically distributed exponential random variables with parameter $\nu$.

Recalling that the expected value of an exponential with parameter $\nu$ 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.

The exponential distribution and independence properties of the interarrival times follow directly from the conditions (P1) and (P2) defining the Poisson process. Consider the first arrival time $T_1$. 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 (5.27), $T_1$ is exponential with parameter $\nu$.

Now, let’s restart the counting process at time $T_1$ by defining

$$Q_t \triangleq N_{T_1 + t} - N_{T_1}.$$ 

$Q_t$ restarts the counting of arrivals at time $T_1$; it is the increment of $N$ over the interval $(T_1, T_1 + t]$. It can be shown that conditions (P1) and (P2) of the definition of Poisson processes apply to this increment, even though $T_1$ is random. (This is intuitively reasonable, but not simple to prove rigorously.) Thus, applying (P1), $Q_t$ is a Poisson random variable with parameter $\nu t$. Reasoning as for $T_1$,

$$\mathbb{P}(T_2 - T_1 > t) = \mathbb{P}(Q_t = 0) = e^{-\nu t}$$
for all \( t \) and so \( T_2 - T_1 \) is exponential. Applying condition (P2), \( \{Q_t\} \) is independent of what happens up to time \( T_1 \). Thus since \( T_2 - T_1 \) depends only on \( \{Q_t\} \), it also is independent of \( T_1 \). Thus we have shown (non-rigorously!) that \( T_1 \) and \( T_2 - T_1 \) are i.i.d. exponential random variables.

To show that \( T_3 - T_2 \) is exponential with parameter \( \nu \) and independent of \( T_3 - T_2 \), we look at \( N_{T_2+t} - N_{T_2} \), which restarts the counting at \( T_2 \) and proceed in the same manner. Continuing like this, it follows that \( T_1, T_2 - T_1, T_3 - T_2, \ldots \) are i.i.d. exponential random variables, as claimed in Theorem 4.

**Example 5.6.** Consider the arrival problem of Example 5.4. What is the expected time to the fifth arrival?

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 is exponential with parameter \( \lambda = 20 \), and its expected value is \( 1/20 \). Thus the expected time to arrival of the fifth call is \( 5(1/20) = 0.25 \) hours. This is the answer you would just guess from the fact that the rate of arrivals is 20 per hour. \( \diamond \)

**Example 5.7.** (Application to restriction enzyme digests) The recognition sequence of enzyme \( \textit{BamHI} \) is \( \textit{GGATCC} \). 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 \( \textit{BamHI} \). 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 \( \textit{BamHI} \) 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 segment, 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 \( \textit{BamHI} \), so the "time" \( T_i \) of arrival \( i \) is just the site of the \( i^{\text{th}} \) 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 $\lambda$ is $1/\lambda$, the average fragment length is

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

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 $1/p$. \hfill \diamond

### 5.3.5 Poisson thinning

Let $N_t$ be a Poisson process with rate $\nu$. 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 $\mu$ 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 $\mu$, 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 $\nu$. Let $\{M_t\}$ be obtained by thinning $\{N_t\}$. If $\mu$ 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 $\nu$ 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 everytime we toss the time, if it comes up heads we then turn over the coin to get tails with probability $1 - \mu$ and leave it heads with probability $\mu$. 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 5.8. Partial Digests.** Suppose *BamHI* (see Example 5.7) 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 segment. 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 total 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 $1/\lambda = 2^{13} = 8192$ base pairs.
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 \( \nu_1 \), counting one stream, and a second Poisson process \( Q_t, t \geq 0 \), with rate \( \nu_2 \), 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 \( \nu_1 \) and \( \nu_2 \), 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 \( \nu_1 \), \( M_t - M_s \) 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 5.9. Double digest.** 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 5.24.** 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 5.25.** 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 5.26. 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 \( \Gamma(r) \overset{\Delta}{=} \int_0^\infty x^{r-1} e^{-x} \, dx \). The reason we introduce this random variable is the following result: if \( X \) and \( Y \) are independent exponential random variables each having parameter \( \lambda \), then \( X + Y \) has the gamma distribution with parameters \( \lambda \) 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

\[ M_X(t) = M_Y(t) = \frac{\lambda}{\lambda - t}, \quad \text{if } t < \lambda. \]

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 referring to the table in 2.3.5, this is precisely the moment generating function of a gamma random variable with parameters \( \lambda \) 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 \( \nu \) and consider a time \( t > 0 \). The \textbf{residual lifetime}, \( R_t \), is the time that elapses from \( t \) until the next arrival strictly later than \( t \). The \textbf{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 \( t \), \( C_t = 0 \).)

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

$$P(R_t > s) = P(N(t+s) - N(t) = 0) = e^{-\nu s}. \quad (5.28)$$

This says that the residual lifetime is exponential with parameter $\nu$, 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 $P(C_t > t) = 0$. 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 $P(N_t - N_{t-s} = 0) = e^{-\nu s}$. To summarize,

$$\begin{cases}
P(C_t > s) = 0, & \text{if } s > t; \\
P(C_t \geq s) = e^{-\nu s}, & \text{if } 0 \leq s \leq t.
\end{cases} \quad (5.29)$$

Notice that

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

We see that for $s < t$, this distribution looks precisely like the exponential distribution with parameter $\nu$, 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

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

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 $(t, \infty)$, 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 $\nu$. Since $R_t$ is also exponential with parameter $\nu$ 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 $\nu$ and $r = 2$. This is the main result we wanted to get to.

**Example 5.10** 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 containing 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 $\Gamma(2) = 1$,

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

Upon integration by parts,

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

**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

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

\( I_x \) indicates whether \( x \) is in a sequencable fragment or not. The proportion of the segment covered by sequencable fragments is therefore

\[
C = \frac{1}{g} \int_0^g I_x \, dx,
\]

and the expected coverage is

\[
E[C] = \frac{1}{g} \int_0^g E[I_x] \, dx = \frac{1}{g} \int_0^g P(I_x = 1) \, dx.
\tag{5.30}
\]

To derive a formula for \( E[C] \), we will compute \( P(I_x = 1) \), 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 \( \nu \). 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

\[
f(y) = \nu^2 ye^{-\nu y}, \quad y > 0.
\]

So, using an integration by parts

\[
P(I_x = 1) \approx \int_L^U \nu^2 ye^{-\nu y} \, dy = (\nu L + 1)e^{-\nu L} - (\nu + 1)e^{-\nu U}.
\]

This answer does not depend on \( x \) (ignoring end effects) and substitution into equation (5.30) gives the coverage formula:

\[
E[C] \approx (\nu L + 1)e^{-\nu L} - (\nu U + 1)e^{-\nu U}.
\tag{5.31}
\]
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 $\mu$ 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 $\mu$ be chosen to obtain the best coverage?

We will choose $\mu$ to maximize the expected coverage. According to (5.31), for a digestion probability $\mu$ 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

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

\[\Box\]

5.3.10 Problems

Exercise 5.27. Show that $A(\mu)$ in the last calculation is maximized at $\mu^*$.

Exercise 5.28. 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 5.29. Let $\{N_t\}$ count the number of sites cut by a digest be Poisson with rate $\nu$. 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?