Shotgun sequencing:

Ex. Suppose we have the segment ACTAATGACCCATGG.
To shotgun sequence, we first take lots of duplicates of this segment.
We then chop up the duplicates at random points, and throw these segment fragments together.
We randomly select N of the fragments and sequence them.
We fit the fragments together by looking for overlaps.

ACTAATGACCCATGG
\[ \downarrow \] duplicate segment and break into fragments: we take N=3 of them.
ACTA, CATGG, TAAT
\[ \downarrow \] we look for overlaps, and reconstruct the segment.
ACTAAT \_ \_ CATGG.

The fitted-together fragments are called "contigs?"

\[ \text{contig} \#1 \quad \text{contig} \#2 \quad \text{contig} \#3 \]

We see that the coverage \( C = \frac{\text{total length of all the contigs}}{g} \in [0,1] \)

where \( g \) is the length of the segment to be sequenced.

Model I: Fragments are all of fixed length \( L \).
Fragment \( i \) is located at \([x_i, x_i + L]\) (\( x_i \) is the left endpoint of fragment).
\( x_i \) is independent of \( x_j \) (\( j \neq i \)) and is uniformly distributed on \([0, g] \). \( (P(a < x_i < b) = \frac{b-a}{g}) \)
\( g \) = length of DNA segment.
We assume that \( L \) is very small compared to \( g \),
and that \( N \) (the number of fragments we take) is very large.

We will prove: the Clark-Carbone formula,

\[ E[C] = 1 - e^{-\frac{NL}{g}}. \]
Remember that $E[x] = \int x \cdot f_x(x) \, dx$.

To start, pick a point $y$ s.t. $y \in [0, g]$.

What is the probability that a randomly chosen fragment will cover $y$?

Ignore the boundary problem (when $y$ is very close to 0 or $g$).

For a randomly chosen fragment to cover $y$, the left endpoint $x$ must lie inside $[y-L, y]$.

$$P(y-L < x < y) = \frac{y-(y-L)}{g} = \frac{L}{g} \quad (\text{since } x \text{ is uniform}).$$

We next introduce a random variable called $I$.

$$I(y) = \begin{cases} 
  1 & \text{if at least one fragment covers } y \\
  0 & \text{otherwise}
\end{cases}$$

$$= \begin{cases} 
  1 & \text{if } y \text{ is covered by a contig} \\
  0 & \text{if } y \text{ is in a gap}
\end{cases}$$

$$P(I(y) = 0) = P(\text{no fragments cover } y), \text{ but since the fragments are independent of each other,}$$

$$P(I(y) = 0) = P(\text{fragment 1 does not cover } y) \cdot P(\text{fragment 2 does not cover } y) \cdots P(\text{fragment } N \text{ does not cover } y)$$

$$= \left[ P(\text{a fragment does not cover } y) \right]^N$$

$$= \left(1 - \frac{L}{g}\right)^N$$

Since $I(y)$ is a Bernoulli random variable,

$$P(I(y) = 1) = 1 - P(I(y) = 0)$$

$$= 1 - \left(1 - \frac{L}{g}\right)^N.$$
What is the expected value of $C$?

$$C = \frac{\text{total length of all contigs}}{g} = \frac{1}{3} \int_0^g I(y) \, dy.$$ 

$$E[C] = E\left[\frac{1}{3} \int_0^g I(y) \, dy\right] = \frac{1}{3} \int_0^g E[I(y)] \, dy$$

$$E[I(y)] = O \cdot P(I(y) = 0) + 1 \cdot P(I(y) = 1)$$

$$= 1 - \left(1 - \frac{N}{g}\right)^N.$$ 

$$E[C] = \frac{1}{3} \int_0^g \left(1 - \left(1 - \frac{N}{g}\right)^N\right) \, dy = 1 - \left(1 - \frac{N}{g}\right)^N$$

$$= 1 - \left(1 - \frac{N}{g}\right)^N$$

$$= 1 - \left(1 - \frac{NL}{g}\right)^N$$

$$\leq 1 - e^{-\frac{NL}{g}}.$$ 

What is the expected number of contigs?

Let $W$ be a random variable representing the number of contigs.

Introduce new random variables $Z_i$, $i = 1, 2, \ldots, N$

$$Z_i = \begin{cases} 1 & \text{if fragment } i \text{ is the rightmost fragment of a contig} \\
0 & \text{otherwise} \end{cases}$$

Ex.

$$Z_1 = 0, \quad Z_2 = 1, \quad Z_3 = 0, \quad Z_4 = 1$$

$$Z_5 = 0, \quad Z_6 = 0, \quad Z_7 = 1.$$ 

Notice that $W = \sum_{i=1}^N Z_i = 3$ in this example.

Since $E[W] = E[\sum_{i=1}^N Z_i] = \sum_{i=1}^N E[Z_i]$ and $E[Z_i] = O \cdot P(Z_i = 0) + 1 \cdot P(Z_i = 1)$,

we see that $E[W] = \sum_{i=1}^N P(Z_i = 1)$. 

What is the probability that $Z_i = 1$? Let $i$ be arbitrary.

Claim: The event $\{Z_i = 1\}$ is the same as $\{\text{no } x_j (j \neq i) \text{ falls in } [x_i, x_i + L]\}$.

Example:

\[
P(x_j \text{ does not fall in } [x_i, x_i + L]) = (1 - \frac{L}{g})
\]

\[
P(Z_i = 1) = (1 - \frac{L}{g})^{N-1}
\]

\[
= (1 - \frac{1}{N-1} \cdot \frac{(N-1)L}{g})^{N-1}
\]

Let $a = \frac{L(N-1)}{g} \geq \frac{L \cdot N}{g}$ for large $N$, so

\[
P(Z_i = 1) \leq e^{-a} \left( e^{\frac{LN}{g}} \text{ or } e^{\frac{L \cdot (N-1)}{g}} \right).
\]

Therefore, we conclude that:

\[
E[W] = \sum_{i=1}^{N} P(Z_i = 1) \leq N \cdot e^{-a} \left( N \cdot e^{\frac{-NL}{g}} \text{ or } N \cdot e^{\frac{-(N-1)L}{g}} \right).
\]
Model II: random fragment lengths $L_i$

$X_i$: independent and uniformly distributed (as in Model I).

$L_i$: independent, equally distributed with density $f_L(l)$ and independent of $X$.

We assume that $f_L(l) = 0$ if $L > K$ for some $K < g$.

\[ f_L(l) \]

\[ K \]

\[ g \]

\[ l \]

We will prove: $E[C] = 1 - e^{-\frac{NL}{g}}$.

First, what is the density of the joint distribution of $(X_i, L_j)$?

\[ f_n(x, l) = \text{product of density of } x_i \text{ with density of } l \]

\[ = \begin{cases} \frac{1}{g} \cdot f_L(l) & \text{if } 0 \leq x \leq g \\ 0 & \text{otherwise} \end{cases} \]

This implies that for any set $A$,

\[ P((X_i, L_j) \in A) = \int \int_A \frac{1}{g} f_L(l) \, dx \, dl \]

Now what is the probability that a given point $y_0$ is covered by a fragment?

Pick $y_0 \in [0, g]$.

Let $I(y_0) = \begin{cases} 1 \text{ if at least one fragment covers } y_0 \\ 0 \text{ otherwise} \end{cases}$

\[ E[I(y_0)] = \cdot P(I(y_0) = 0) + 1 \cdot P(I(y_0) = 1) \]

\[ = P(I(y_0) = 1) \]
What is $P(I(y_0) = 1)$? We can see that the event that a particular segment $[x_i, x_i + L_i]$ covers $y_0$ is the same as the event that $x_i \in [y_0 - L_i, y_0]$.

We can also see that:

$$P(I(y_0) = 1) = P(x_i \leq y_0 \text{ and } x_i + L_i \geq y_0)$$

$$P((x_i, L_i) \in A) = \int_A \frac{1}{g} f_L(e) \, dx \, de,$$

where $A$ is the set of $(x, e)$ such that $0 \leq x \leq y_0$, $y_0 \leq x + e$, and $e \leq K$.

What does the area $A$ look like?

$$\int_A \frac{1}{g} f_L(e) \, dx \, de = \int_0^K \left( \int_{y_0 - L}^{y_0} \frac{1}{g} f_L(e) \, dx \right) \, de$$

$$= \int_0^K \frac{1}{g} f_L(e) \cdot (\int_{y_0 - L}^{y_0} dx) \, de$$

$$= \frac{1}{g} \int_0^K e \cdot f_L(e) \, de$$

$$= \frac{1}{g} E[e].$$

Therefore, $P((x, L) \in A) = \frac{1}{g} E[e]$. 

\[ P(\text{fragment 1 does not cover } y_0) = 1 - \frac{E[L]}{g}, \]
\[ P(\text{fragment 2 does not cover } y_0) = 1 - \frac{E[L]}{g}, \]
\[ \vdots \]
\[ P(\text{fragment N does not cover } y_0) = 1 - \frac{E[L]}{g}. \]
\[ P(\text{no fragment covers } y_0) = \left(1 - \frac{E[L]}{g}\right)^N. \]

So we see that:
\[ P(\text{some fragment covers } y_0) = P(I(y_0) = 1) = 1 - \left(1 - \frac{E[L]}{g}\right)^N. \]
\[ E[I(y_0)] = \frac{1}{g} \int_0^g I(y) \, dy, \]
\[ C = \frac{1}{g} \int_0^g I(y) \, dy. \]
\[ E[C] = E\left[ \frac{1}{g} \int_0^g I(y) \, dy \right] \]
\[ = \frac{1}{g} \int_0^g E[I(y)] \, dy \]
\[ = \frac{1}{g} \int_0^g \left(1 - \left(1 - \frac{E[L]}{g}\right)^N \right) \, dy \]
\[ = 1 - \left(1 - \frac{E[L]}{g}\right)^N \]
\[ \leq 1 - e^{\frac{NE[L]}{g}}, \text{ as we've seen before}. \]

We conclude that:
\[ E[C] \leq 1 - e^{\frac{NE[L]}{g}}. \]