Binomial Distribution

• **Binomially-distributed** random variable $X$ equals sum (number of successes) of $n$ independent Bernoulli trials

• The probability mass function is:

\[ f(x) = C_x^n p^x (1 - p)^{n-x} \text{ for } x = 0, 1, \ldots n \quad (3-7) \]

• Based on the binomial expansion:

\[ (p + q)^n = \sum_{x=0}^{n} C_x^n p^x q^{n-x} \]
Poisson Distribution

• Limit of the binomial distribution when
  – \( n \), the number of attempts, is very large
  – \( p \), the probability of success is very small
  – \( E(X)=np \) is just right

The annual numbers of deaths from horse kicks in 14 Prussian army corps between 1875 and 1894

<table>
<thead>
<tr>
<th>Number of deaths</th>
<th>Observed frequency</th>
<th>Expected frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>144</td>
<td>139</td>
</tr>
<tr>
<td>1</td>
<td>91</td>
<td>97</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5 and over</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>280</td>
<td>280</td>
</tr>
</tbody>
</table>

From von Bortkiewicz 1898
Let $\lambda = np = E(x)$, so $p = \frac{\lambda}{n}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \frac{n(n-1)\ldots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \sim \frac{n^x}{x!} \left(\frac{\lambda}{n}\right)^x = \frac{\lambda^x}{x!};$$

$$\sum_x \frac{\lambda^x}{x!} = e^\lambda.$$

Normalization requires $\sum_x P(X = x) = 1$.

Thus $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$
Poisson Mean & Variance

If $X$ is a Poisson random variable, then:

• **Mean:** $\mu = E(X) = \lambda = \nu \cdot \rho$

• **Variance:** $\sigma^2 = V(X) = \lambda = \nu \cdot \rho \cdot (1 - \rho) \approx \nu \cdot \rho$

• **Standard deviation:** $\sigma = \lambda^{1/2}$

Note: Variance = Mean

Note: Standard deviation/Mean = $\lambda^{-1/2}$

decreases with $\lambda$
Poisson Distribution in Genome Assembly
Poisson Example: Genome Assembly

- **Goal**: figure out the sequence of DNA nucleotides (ACTG) along the entire genome
- **Problem**: Sequencers generate random short reads

**Table 9.1** Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies’ websites, http://en.wikipedia.org/wiki/DNA_sequencer, and literature cited for each technology.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
<th>Reads per run</th>
<th>Time per run</th>
<th>Cost per megabase (US$)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>700</td>
<td>1 million</td>
<td>1 day</td>
<td>10</td>
<td>99.90</td>
</tr>
<tr>
<td>Illumina</td>
<td>50–250</td>
<td>&lt;3 billion</td>
<td>1–10 days</td>
<td>~0.10</td>
<td>98</td>
</tr>
<tr>
<td>SOLiD</td>
<td>50</td>
<td>~1.4 billion</td>
<td>7–14 days</td>
<td>0.13</td>
<td>99.90</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>200</td>
<td>&lt;5 million</td>
<td>2 hours</td>
<td>1</td>
<td>98</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>2900</td>
<td>&lt;75,000</td>
<td>&lt;2 hours</td>
<td>2</td>
<td>99</td>
</tr>
<tr>
<td>Sanger</td>
<td>400–900</td>
<td>N/A</td>
<td>&lt;3 hours</td>
<td>2400</td>
<td>99.90</td>
</tr>
</tbody>
</table>

- **Solution**: assemble genome from short reads using computers. **Whole Genome Shotgun Assembly**.
Genome Assembly

Whole-genome “shotgun” sequencing starts by copying and fragmenting the DNA

(“Shotgun” refers to the random fragmentation of the whole genome; like it was fired from a shotgun)

Input:  GGCCTCTATATCTCGGCCTCTAGGCCCCTCATTTTTTT

35bp

Copy by PCR:  GGCCTCTATATCTCGGCCTCTAGGCCCCTCATTTTTTT

Fragment:  GGCCTCTA   TATCTCGG   CTCTAGGCCCCTC   ATTTTTTT
          GGC    GTCTATAT   CTCGGCTCTAGGCCCCTCA    TTTTTTT
          GGCCTC   TATATCT   CGGCTCTAGGCCCCT    CATTTTTTT
          GGCCTCTAT   ATCTCGGCTCTAG    GCCCTCA   TTTTTTT

Courtesy of Ben Langmead. Used with permission.

http://www.langmead-lab.org/teaching-materials/
Assembly

Assume sequencing produces such a large number of fragments that almost all genome positions are covered by many fragments...

...but we don’t know what came from where

Reconstruct this

From these

Courtesy of Ben Langmead. Used with permission.

http://www.langmead-lab.org/teaching-materials/
Assembly

Overlaps between short reads help to put them together

CTAGGCCCTCAATTTTTT 177 nucleotides
CTCTAGGCCCTCAATTTTTT
GGCTCTAGGCCCTCATTTTTT
CTCGGCTCTAGCCCCTCATTTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCTATATCT
GGCGTCTATATCT
GGCGTCTATATCTCG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCTATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

35 nucleotides

Courtesy of Ben Langmead. Used with permission.

http://www.langmead-lab.org/teaching-materials/
Short Reads assemble into Contigs

Figure 5.1.
How many short reads do we need?

Input

Low coverage:

A few pieces to assemble

High coverage:

many pieces to assemble

Output

many contigs, many gaps

a few contigs, a few gaps
Where is the Poisson?

- $G$ - genome length (in bp)
- $L$ - short read average length
- $N$ – number of short read sequenced
- $\lambda$ – sequencing coverage redundancy = $LN/G$
- $x$ - number of short reads covering a given site on the genome

\[
P(x) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): $p=L/G$ is very small. Number of attempts (short reads): $N$ is very large. Their product (sequencing redundancy): $\lambda = NL/G$ is $O(1)$.

Ewens, Grant, Chapter 5.1
What fraction of the genome is missing?
What fraction of genome is covered?

- Coverage: $\lambda = NL/G$,
  
  $X$ – random variable equal to the number of times a given site is covered by short reads.

  Poisson: $P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}$

  $P(X = 0) = \exp(-\lambda)$, $P(X > 0) = 1 - \exp(-\lambda)$

- Total length covered: $G \times [1 - \exp(-\lambda)]$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean proportion of genome covered</td>
<td>.864665</td>
<td>.981684</td>
<td>.997521</td>
<td>.999665</td>
<td>.999955</td>
<td>.999994</td>
</tr>
</tbody>
</table>

Table 5.1. The mean proportion of the genome covered for different values of $\lambda$.
How many contigs?
How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within it.
- The left end of another short read has the probability \( p = \frac{L-1}{G} \) to fall within a given read. There are \( N-1 \) other reads. Hence the expected number of left ends inside a given shot read is \( p \cdot (N-1) = \frac{(N-1) \cdot (L-1)}{G} \approx \lambda \).
- If significant overlap required to merge two short reads is \( L_{ov} \), modified \( \lambda \) is given by \( (N-1) \cdot (L - L_{ov})/G \).
- Probability that no left ends fall inside a short read is \( \exp(-\lambda) \). Thus the Number of contigs is \( N_{contigs} = N \exp(-\lambda) \).

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean number of contigs</td>
<td>60.7</td>
<td>70.8</td>
<td>73.6</td>
<td>66.9</td>
<td>54.1</td>
<td>29.9</td>
<td>14.7</td>
<td>6.7</td>
<td>3.0</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Table 5.2. The mean number of contigs for different levels of coverage, with \( G = 100,000 \) and \( L = 500 \).
Average length of a contig?

- Length of a genome covered:
  \[ G_{\text{covered}} = G \cdot P(X>0) = G \cdot (1 - \exp(-\lambda)) \]
- Number of contigs \( N_{\text{contigs}} = N \cdot e^{-\lambda} \)
- Average length of a contig =
  \[ <L> = \frac{\sum_i L_i}{N_{\text{contigs}}} = \frac{G_{\text{covered}}}{N_{\text{contigs}}} = G \cdot (1 - \exp(-\lambda))/N \cdot e^{-\lambda} = L \cdot (1 - \exp(-\lambda))/\lambda \cdot e^{-\lambda} \]

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean contig size</td>
<td>1,600</td>
<td>6,700</td>
<td>33,500</td>
<td>186,000</td>
<td>1,100,000</td>
</tr>
</tbody>
</table>

Table 5.3. The mean contig size for different values of \( \lambda \) for the case \( L = 500 \).
Estimate

• Human genome is $3 \times 10^9$ bp long
• Chromosome 1 is about $G = 0.25 \times 10^9$ bp
• Illumina generates short reads $L = 100$ bp long
• What number of reads $N$ are needed to completely assemble the 1st chromosome?
  • The formula to use is: $1 = N_{contigs} = Ne^{-\lambda} = Ne^{-NL/G}$
  • Answer: $N = 4.4 \times 10^7$ short (100bp) reads
  Test: $4.4e7*exp(-4.4e7*100/0.25e9)=0.9997$
  • What coverage redundancy $\lambda$ will it be?
  Answer: $\lambda = NL/G = 17.6$ coverage redundancy
How much would it cost to assemble human genome now?

- Human Genome Project: $2.7 billion in 1991 dollars.
- Now a **de novo full assembly** of the whole human genome would now cost $3 \times 10^9 \times 17.6 / 10^6 \times 0.1$/MB = $5300
- 2\textsuperscript{nd} genome (and after) would be even cheaper as we would already have a reference genome to which we can map short reads. (Puzzle: picture on the box)
- But, this is a naïve estimate. In reality there are complications. See next slides:
What spoils these estimates?

>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic contig, GRCh37.p13 Primary Assembly (displaying 3’ end)
CGGGAAATCAAAAAAGCCCTCTGGAATCTGCGCAACCAGATTCTCCCAAGCAAAGTGAGCGGCGAGCGT
GGGAGATCCACACCGTAGCATTTGAACACAAATGCAACATTACAAATGACACATGACACCGAAATATA
ACACACCCCATGCTCTGTAACAAAGCACCCTGTAATGCTAATGCACTGCTCTAAACAAATATAATAT
AAGATCGGCAATCCCGACACTGCGGTCAGTCTAAGACAGCAATGAAAATAGTCAACATAATAACCTGA
ATAGTGTAGGGTTAGGGTCAGGGTCCCGGTCCCGGTCGCGGTTCGGGGTCGCGGTTCGGGGGCAGGGTGAG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT

FIGURE 8.11  A BLASTN search of the human genome (all assemblies) database was performed at
the NCBI website using **TTAGGGTTAGGGTTAGGG** as query (i.e., three TTAGGG repeats). There
were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned
to the telomere of chromosome 12q having many dozens of TTAGGG repeats. These occurred at the 3’
end of the genomic contig sequence.

There were **100s of matches** while one expects **<< 1 match:**

\[2 \cdot 3 \times 10^9 \cdot 4^{-18} = 0.08 << 1\]

DNA repeats make assembly difficult
Repeats are like sky puzzle pieces
How many repeats are in eukaryotic genomes?

Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves

**FIGURE 8.6** The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a $C_0 t_{1/2}$ curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0 t_{1/2}$ value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow A), and represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.

© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics