A gallery of useful discrete probability distributions
Bernoulli distribution

The simplest non-uniform distribution

\[ p \text{ – probability of success (1)} \]
\[ 1-p \text{ – probability of failure (0)} \]

\[
f(x) = P(X = x) = \begin{cases} 
  p & \text{if } x = 1 \\
  1-p & \text{if } x = 0 
\end{cases}
\]

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant \( e=2.718... \)
Bernoulli distribution

\[ f(x) = P(X = x) = \begin{cases} 
  p & \text{if } x = 1 \\
  1 - p & \text{if } x = 0 
\end{cases} \]

\[ E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p \]

\[ \text{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p) \]
Refresher: Binomial Coefficients

\[
\binom{n}{k} = C^n_k = \frac{n!}{k!(n-k)!}, \text{ called } n \text{ choose } k
\]

\[
\binom{10}{3} = C^{10}_3 = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3 \cdot 2 \cdot 1 \cdot 7!} = 120, \text{ called 10 choose 3}
\]

Number of ways to choose \( k \) objects out of \( n \) without replacement and where the order does not matter. Called binomial coefficients because of the binomial formula

\[
(p + q)^n = \sum_{x=0}^{n} C^n_x p^x q^{n-x}
\]
Binomial Distribution

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

• The probability mass function is:

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

• Based on the binomial expansion:

$$\sum_{x=0}^{n} C_x^n p^x q^{n-x} = (p+q)^n \quad (3-7)$$
Binomial Mean and Variance

$X$ is a binomial random variable with parameters $p$ and $n$.

Mean:
$\mu = E(X) = np$

Variance:
$\sigma^2 = V(X) = np(1-p)$

Standard deviation:
$\sigma = \sqrt{np(1-p)}$
\[ q = 1 - p \Rightarrow p + q = 1 \Rightarrow (p + q)^n = 1 \]

\[
(p + q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = 1
\]

\[ E(X) = \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x q^{n-x} = p \cdot \frac{\partial}{\partial p} (p + q)^n \bigg|_{p + q = 1}
\]

\[ = p \cdot n \]

Some trick:

\[ E(x(x-1)) = p^2 \frac{\partial}{\partial p^2} (p + q)^n = p^2 n(n-1) \]

\[ E(x^2) = E(x(x-1)) + E(x) = p^2 n(n-1) + pn \]

\[ (E(x))^2 = p^2 n^2 \Rightarrow \text{Var}(X) = pn - p^2 n = np(1-p) \]
Matlab exercise: Binomial distribution

• Generate a sample of size 100,000 for binomially-distributed random variable X with n=100, p=0.2
• Tip: generate n Bernoulli random variables and use sum to add them up
• Plot the approximation to the Probability Mass Function based on this sample
• Calculate the mean and variance of this sample and compare it to theoretical calculations: E[X]=n*p and V[X]=n*p*(1-p)
Matlab exercise: Binomial distribution

- n=100; p=0.2;
- Stats=100000;
- r1=rand(Stats,n)<p;
- r2=sum(r1,2);
- mean(r2)
- var(r2)
- [a,b]=hist(r2, 0:n);
- p_b=a./sum(a);
- figure; stem(b,p_b);
- figure; semilogy(b,p_b,'ko-')
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

Siméon Denis Poisson (1781–1840)
French mathematician and physicist
Let $\lambda = np = E(x)$, so $p = \lambda/n$

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

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

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

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**: $\mu = E(X) = \lambda = \eta \cdot \rho$
- **Variance**: $\sigma^2 = V(X) = \lambda \approx \eta \cdot \rho \cdot (1 - \rho) \approx \eta \cdot \rho$
- **Standard deviation**: $\sigma = \lambda^{1/2}$

Note: Variance = Mean
Note: Standard deviation/Mean = $\lambda^{-1/2}$ decreases with $\lambda$
Poisson Distribution in Genome Assembly
Cost per Megabase of DNA Sequence

Moore's Law

National Human Genome Research Institute
genome.gov/sequencingcosts
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.

<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.**
MinION, a palm-sized gene sequencer made by UK-based Oxford Nanopore Technologies
Short Reads assemble into Contigs

Figure 5.1.