Discrete Probability Distributions
Random Variables

• A variable that associates a number with the outcome of a random experiment is called a random variable.

• Notation: random variable is denoted by an uppercase letter, such as $X$. After the experiment is conducted, the measured value is denoted by a lowercase letter, such as $x$. Both $X$ and $x$ are shown in italics, e.g., $P(X=x)$. 
Continuous & Discrete Random Variables

- A **discrete random variable** is usually integer number
  - $N$ - the number of p53 proteins in a cell
  - $D$ - the number of nucleotides different between two sequences

- A **continuous random variable** is a real number
  - $C=N/V$ – the concentration of p53 protein in a cell of volume $V$
  - Percentage $(D/L) \times 100\%$ of different nucleotides in protein sequences of different lengths $L$
    (depending on the set of $L$’s may be discrete but dense)
Probability Mass Function (PMF)

• I want to compare all 4-mers in a pair of human genomes

• $X$ – random variable: the number of nucleotide differences in a given 4-mer

• Probability Mass Function: $f(x)$ or $P(X=x)$ – the probability that the # of SNPs is exactly equal to $x$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6561</td>
</tr>
<tr>
<td>1</td>
<td>0.2916</td>
</tr>
<tr>
<td>2</td>
<td>0.0486</td>
</tr>
<tr>
<td>3</td>
<td>0.0036</td>
</tr>
<tr>
<td>4</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

$\sum x P(X=x) = 1.0000$
Cumulative Distribution Function (CDF)

### Cumulative Distribution Function CDF:
\[ F(x) = P(X \leq x) \]

**Example:**
\[ F(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999 \]

### Complementary Cumulative Distribution Function (tail distribution) or CCDF:
\[ F_>(x) = P(X > x) \]

**Example:**
\[ F_>(3) = P(X > 3) = 1 - P(X \leq 3) = 0.0001 \]
Mean or Expected Value of $X$

The mean or expected value of the discrete random variable $X$, denoted as $\mu$ or $E(X)$, is

$$\mu = E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)$$

- The mean = the weighted average of all possible values of $X$. It represents its “center of mass”

- The mean may, or may not, be an allowed value of $X$

- It is also called the arithmetic mean (to distinguish from e.g., the geometric mean discussed later)

- Mean may be infinite if $X$ any positive integer and the tail $P(X=x)\geq c/x^2$
Variance $V(X)$: square of a typical deviation from the mean $\mu = E(X)$

$V(X) = \sigma^2$, where $\sigma$ is called standard deviation.

\[ \sigma^2 = V(X) = E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - (E(X))^2 \]
Variance of a Random Variable

If $X$ is a discrete random variable with probability mass function $f(x)$,

$$E[h(X)] = \sum_x h(x) \cdot P(X = x) = \sum_x h(x)f(x) \quad (3-4)$$

If $h(X) = (X - \mu)^2$, then its expectation, $V(X)$, is the variance of $X$. $\sigma = \sqrt{V(X)}$, is called standard deviation of $X$.

$$\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x) \text{ is the definitional formula}$$

$$= \sum_x (x^2 - 2\mu x + \mu^2) f(x)$$

$$= \sum_x x^2 f(x) - 2\mu \sum_x xf(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2$$

$$= \sum_x x^2 f(x) - \mu^2 \text{ is the computational formula}$$

Variance can be infinite if $X$ can be any positive integer and tail of $P(X=x) \geq c/x^3$
Skewness of a random variable

• Want to quantify how asymmetric is the distribution around the mean?
• Need any odd moment: $E[(X-\mu)^{2n+1}]$
• Cannot do it with the first moment: $E[X-\mu]=0$
• Normalized 3-rd moment is skewness: $\gamma_1 = E[(X-\mu)^3]/\sigma^3$
• Skewness can be infinite if $X$ takes unbounded positive integer values and the tail $P(X=x) \geq c/x^4$ for large $x$
Geometric mean of a random variable

• Useful for very broad distributions (many orders of magnitude)?
• Mean may be dominated by very unlikely but very large events. Think of a lottery
• Exponent of the mean of $\log X$: $\text{Geometric mean}=\exp(\mathbb{E}[\log X])$
• Geometric mean usually is not infinite
Summary: Parameters of a Probability Distribution

- **Probability Mass Function (PMF):** \( f(x) = \text{Prob}(X=x) \)
- **Cumulative Distribution Function (CDF):** \( F(x) = \text{Prob}(X \leq x) \)
- **Complementary Cumulative Distribution Function (CCDF):** \( F_>(x) = \text{Prob}(X > x) \)

- **The mean**, \( \mu = E[X] \), is a measure of the center of mass of a random variable.

- **The variance**, \( V(X) = E[(X - \mu)^2] \), is a measure of the dispersion of a random variable around its mean.

- **The standard deviation**, \( \sigma = [V(X)]^{1/2} \), is another measure of the dispersion around mean. Has the same units as \( X \).

- **The skewness**, \( \gamma_1 = E[(X-\mu)^3/\sigma^3] \), a measure of asymmetry around mean.

- **The geometric mean**, \( \exp(E[\log X]) \) is useful for very broad distributions.
Skewness of a random variable

• Want to quantify how asymmetric is the distribution around the mean?
• Need any odd moment: $E[(X-\mu)^{2n+1}]$
• Cannot do it with the first moment: $E[X-\mu]=0$
• Normalized 3-rd moment is skewness: $\gamma_1=E[(X-\mu)^3/\sigma^3]$
• Skewness can be infinite if $X$ takes unbounded integer values and tail $P(X=x) \geq c/x^4$
Geometric mean of a random variable

• Useful for very broad distributions (many orders of magnitude)?
• Mean may be dominated by very unlikely but very large events. Think of a lottery
• Exponent of the mean of $\log X$: Geometric mean=$\exp(E[\log X])$
• Geometric mean usually is not infinite
Summary: Parameters of a Probability Distribution

- **Probability Mass Function (PMF):** $f(x) = \text{Prob}(X=x)$
- **Cumulative Distribution Function (CDF):** $F(x) = \text{Prob}(X \leq x)$
- **Complementary Cumulative Distribution Function (CCDF):** $F_>(x) = \text{Prob}(X > x)$
- The **mean**, $\mu = E[X]$, is a measure of the center of mass of a random variable
- The **variance**, $V(X) = E[(X - \mu)^2]$, is a measure of the dispersion of a random variable around its mean
- The **standard deviation**, $\sigma = [V(X)]^{1/2}$, is another measure of the dispersion around mean. Has the same units as $X$
- The **skewness**, $\gamma_1 = E[(X-\mu)^3/\sigma^3]$, a measure of asymmetry around mean
- The **geometric mean**, $\exp(E[\log X])$ is useful for very broad distributions
A gallery of useful

discrete probability distributions
Discrete Uniform Distribution

• Simplest discrete distribution.
• The random variable $X$ assumes only a finite number of values, each with equal probability.
• A random variable $X$ has a discrete uniform distribution if each of the $n$ values in its range, say $x_1, x_2, \ldots, x_n$, has equal probability.

$$f(x_i) = 1/n$$
Uniform Distribution of Consecutive Integers

• Let $X$ be a discrete uniform random variable all integers from $a$ to $b$ (inclusive). There are $b - a + 1$ integers. Therefore each one gets:

$$f(x) = 1/(b-a+1)$$

• Its measures are:

$$\mu = E(x) = (b+a)/2$$

$$\sigma^2 = V(x) = [(b-a+1)^2-1]/12$$

Note that the mean is the midpoint of $a$ & $b$. 
A random variable $X$ has the same probability for integer numbers $x = 1:10$

What is the behavior of its Probability Mass Function (PMF): $P(X=x)$?

A. does not change with $x=1:10$
B. linearly increases with $x=1:10$
C. linearly decreases with $x=1:10$
D. is a quadratic function of $x=1:10$

Get your i-clickers
A random variable $X$ has the same probability for integer numbers $x = 1:10$

What is the behavior of its Cumulative Distribution Function (CDF): $P(X \leq x)$?

A. does not change with $x = 1:10$
B. linearly increases with $x = 1:10$
C. linearly decreases with $x = 1:10$
D. is a quadratic function of $x = 1:10$

Get your i-clickers
A random variable $X$ has the same probability for integer numbers $x = 1:10$

What is its mean value?

A. 0.5

B. 5.5

C. 5

D. 0.1

Get your i-clickers
A random variable $X$ has the same probability for integer numbers $x = 1:10$

What is its skewness?

A. 0.5  
B. 1  
C. 0  
D. 0.1

Get your i-clickers
An example of the uniform distribution

Cycle threshold (Ct) value in COVID-19 infection
What is the Ct value of a PCR test?

\[ \text{Ct} = \text{const} - \log_2(\text{viral DNA concentration}) \]
Why $C_t$ distribution should it be uniform?

$10^8$

$10^3$

detectable at $C_t \geq 40$

$C_t = 15 - 25$

Random samples from test

$PMF(C_t)$

5 days

time

15 40
Examples of uniform distribution: Ct value of PCR test of a virus

Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledeboer, PhD

American Journal of Clinical Pathology, Volume 154, Issue 4, 1 October 2020,
Why should we care?

- High Ct value means we identified the infected individual early, hopefully before transmission to others.

- When testing is mandatory, and people are tested frequently – Ct value is skewed towards high values.

Non-mandatory tests in Israel

Mandatory tests at UIUC 2021
Matlab exercise: Uniform distribution

• Generate a **sample of size 100,000** for uniform random variable $X$ taking values 1,2,3,...10
• Plot the **approximation** to the probability mass function based on this sample
• Calculate mean and variance of this sample and compare it to infinite sample predictions: $E[X] = (a+b)/2$ and $V[X] = ((a-b+1)^2-1)/12$
Matlab template: Uniform distribution

- b=10; a=1; % b = upper bound; a = lower bound (inclusive)
- Stats=100000; % sample size to generate
- r1=rand(Stats,1);
- r2=floor(???r1)+??;
- mean(r2)
- var(r2)
- std(r2)
- [hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10
  - % hy - number of counts in each bin; hx - coordinates of bins
- p_f=hy./??; % normalize counts to add up to 1
- figure; plot(??,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF
Matlab exercise: Uniform distribution

- b=10; a=1; % b = upper bound; a = lower bound (inclusive)
- Stats=100000; % sample size to generate
- r1=rand(Stats,1);
- r2=floor(b*r1)+a;
- mean(r2)
- var(r2)
- std(r2)
- [hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10
- % hy - number of counts in each bin; hx - coordinates of bins
- p_f=hy./sum(hy); % normalize counts to add up to 1
- figure; plot(hx,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF
Bernoulli distribution

The simplest non-uniform distribution

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

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

- Law of large numbers
- Mathematical constant \( e = 2.718 \ldots \)
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

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

\[ \binom{10}{3} = \binom{10}{3} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3 \cdot 2 \cdot 1 \cdot 7!} = 120 \]

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

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

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

- The probability mass function is:

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

- Based on the binomial expansion:

  $$1 = (p + q)^n = \sum_{x=0}^{n} C^n_x \, p^x \, q^{n-x}$$

Sec 3-6 Binomial Distribution
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:
$\sigma = \sqrt{np(1-p)}$
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 template: Binomial distribution

- \( n=100; \ p=0.2; \)
- \( \text{Stats}=100000; \)
- \( r1=\text{rand}(\text{Stats},n) \ < \text{or} \ > \ p; \)
- \( r2=\text{sum}(r1, \ 1 \text{ rows or 2 columns} \ p); \)
- \( \text{mean}(r2) \)
- \( \text{var}(r2) \)
- \( [a,b]=\text{hist}(r2, 0:n); \)
- \( p_b=\/.\text{sum}(\); \)
- \( \text{figure; stem}(?,p_b); \)
- \( \text{figure; semilogy}(?,p_b,'ko-') \)
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-')
Secretary problem

• An employer has a known and large number – n – of applicants for a secretary position, whom are interviewed one at a time
• Employer can easily evaluate and rank applicants relative to each other but has no idea of the overall distribution of their quality
• Employer has only one chance to choose the secretary, gives yes/no answer in the end of each interview and cannot go back to rejected applicants
• How can employer maximize the probability to choose the best secretary among n applicants?
Martin Gardner (1914 – 2010) Described the *secretary problem* in Scientific American 1960. was an American popular mathematics and popular science writer. Best known for “recreational mathematics”: He was behind the “Mathematical Games” section in Scientific American.

Eugene Dynkin (1924 – 2014) solved this problem in 1963. He referred to it as “picky bride” problem was a Russian and American mathematician, member of NAS. He has made contributions to the fields of probability and algebra. The Dynkin diagram, the Dynkin system, and Dynkin's lemma are all named after him.
What should the employer do?

• Employer does not know the distribution of the quality of applicants and has to learn it on the fly
• Algorithm: look at the first r applicants, remember the best among them
• Hire the first among next n-r applicants who is better than the best among the first r applicants
• How to choose r?
• r small – not enough information: the best among r is not very good. You are likely to hire a looser.
• r=n-1 – you procrastinated too long! You have almost all information but you will have to hire the last applicant who is (likely) not particularly good
\[ P(r) = \sum_{i=1}^{n} P(\text{applicant } i \text{ is selected} \cap \text{applicant } i \text{ is the best}) \]

\[ = \sum_{i=1}^{n} P(\text{applicant } i \text{ is selected} | \text{applicant } i \text{ is the best}) \times P(\text{applicant } i \text{ is the best}) \]

\[ = \left[ \sum_{i=1}^{r-1} 0 + \sum_{i=r}^{n} P\left( \begin{array}{c} \text{the best of the first } i-1 \text{ applicants} \text{ is in the first } r-1 \text{ applicants} \\ \text{applicant } i \text{ is the best} \end{array} \right) \right] \times \frac{1}{n} \]

\[ = \sum_{i=r}^{n} \frac{r-1}{i-1} \times \frac{1}{n} = \frac{r-1}{n} \sum_{i=r}^{n} \frac{1}{i-1}. \]
\[ P(r) = \frac{r - 1}{n} \sum_{i=r}^{n} \frac{1}{i - 1}. \]

Letting \( n \) tend to infinity, writing \( x \) as the limit of \( r/n \), using \( t \) for \( i/n \) and \( dt \) for \( 1/n \),

\[ P(x) = x \int_{x}^{1} \frac{1}{t} dt = -x \ln(x). \]

\[ \frac{dP(x)}{dx} = -\ln(x) - 1 \]

\[ -\ln(x^*) - 1 = 0 \]

\[ x^* = \frac{1}{e} = 0.3679 \]

Probability of picking the best applicant is also \( \frac{1}{e} = 0.3679 \)
Bonus matlab exercise: Picky bride

• Generate a sample of size 100 of applicants with quality given by random number between 0 and 1
• Let the employer select the first applicant exceeding the best applicant among
  – the first 10 applicants (not enough information)
  – the first 37 applicants (optimal)
  – the first 90 applicants (too much procrastination)
• Repeat for 100,000 employers and calculate the probability that employer picked the best applicant in each case
• Test: the probability of picking the best applicant after interviewing 37 applicants the answer should be around \(1/e = 0.37\). In all other cases - smaller
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>

Siméon Denis Poisson (1781–1840)
French mathematician and physicist

From von Bortkiewicz 1898
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: $\mu = E(X) = \lambda \approx n \cdot p$
• Variance: $\sigma^2 = V(X) = \lambda \approx n \cdot p \cdot (1 - p) \approx n \cdot p$
• Standard deviation: $\sigma = \lambda^{1/2}$

Note: Variance = Mean
Note: Standard deviation/Mean = $\lambda^{-1/2}$ decreases with $\lambda$
Matlab exercise: Poisson distribution

• Generate a sample of size 100,000 for Poisson-distributed random variable X with $\lambda = 2$
• 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:
  $E[X] = \lambda$ and $V[X] = \lambda$
Matlab exercise: Poisson distribution

• Stats=100000; lambda=2;
• r2=random('Poisson',lambda,Stats,1);
• mean(r2)
• var(r2)
• [a,b]=hist(r2, 0:max(r2));
• p_p=a./sum(a);
• figure; stem(b,p_p);
• figure; semilogy(b,p_p,'ko-')
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

<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>Error (%)</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>0.1</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>2</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>0.1</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>2</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>1</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>0.1</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.
I think I found the corner piece!
How many short reads do we need?

**Input**
- Low coverage:
  - A few pieces to assemble

**Output**
- many contigs, many gaps

**Input**
- High coverage:
  - many pieces to assemble

**Output**
- 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 redundancy = \( LN/G \)
- \( x \) - number of short reads covering a given site on the genome

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

Ewens, Grant, Chapter 5.1

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 \( O(1) \).
What fraction of genome is covered?

- Coverage: \( \lambda = \frac{N L}{G} \),
  \( X - r.v. \equal{\text{to the number of times a given site is covered}} \)
  Poisson: \( P(X=x) = \frac{\lambda^x \cdot \exp(-\lambda)}{x!} \)
  \( P(X=0) = \exp(-\lambda), \ P(X>0) = 1 - \exp(-\lambda) \)

- Total length covered: \( G \cdot \left[ 1 - \exp(-\lambda) \right] \)

<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?

- Probability that a given short read is the right end of a contig = no left ends of other reads fall within it.
- Left ends of each of \( N-1 \approx N \) other reads has Prob: \( p = (L-1)/N \approx L/N \) to fall within given read. Probability that none do is = \( \exp(-\lambda) \):

\[
N_{\text{contigs}} = N e^{-\lambda}
\]

<table>
<thead>
<tr>
<th>( a )</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 \( a \) for the case \( L = 500 \).
Estimate

- Human genome is $3 \times 10^9$ bp long
- Chromosome 1 spans about $250 \times 10^6$ bp
- Illumina generates short reads 100 bp long
- How many short reads are needed to completely assemble the 1st chromosome?
The answer is

\[ N = 44 \times 10^6 \text{ short (100bp) reads or } \lambda = 17.6 \text{ fold redundant coverage.} \]

At \$0.1/Mb\ that means that the reads for de novo full assembly of human genome would cost

\[ (3 \times 10^9 \times 17.6 / 10^6) \times 0.1 \text{\$} \]

= \$5300 / genome

In reality is cheaper as we don’t need de novo assembly.
What spoils these estimates?

>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic contig, GRCh37.p13 Primary Assembly (displaying 3’ end)
CGGAAATCAAAAGCCCCCTCTGAATCCTGCGACCAGAGATTTCTCCCAGCAAGGTTAGGGCGGCAGCATG
GGGAGATCCACACCGGTAGCTGGACACAAATGCAGCATTTAACATGCAGCATGCACACCCGAAAATATA
ACACACCCATGCTCATGTAACAACGGACCTGTAATGCTAATGCACTGTCCCTAAAACAAATATTAATAT
AAGATCGGCAATCCGCCACACTGCCCCTGCAAGTCTAGAAGCAAGTAAGTACGTCAACATAATAACCCT
ATAGTGGTTAGGCTAAGGCTCCCGGTCGAGGCTCCGGGCTCCGGGTTCAGGGTCA
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
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:

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

\[ \log(2 \cdot 3 \times 10^9)/\log(4) = 16.2412 \] meaning 17 bases should be enough to specify
unique pohuman genome at a.

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

**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
Next 4 slides are based on “Working with Molecular Genetics” course taught by Ross Hardison at Penn State U.

http://www.bx.psu.edu/~ross/workmg/WorkingWithMolecularGeneticsRCHwCover.pdf
Distinct components in complex genomes

• Highly repeated DNA
  – $R$ (repetition frequency) $\geq 100,000$
  – Almost no information, low complexity

• Moderately repeated DNA
  – $10 < R < 10,000$
  – Little information, moderate complexity

• “Single copy” DNA
  – $R=1$ or $2$
  – Much information, high complexity
Almost all transposable elements in mammals fall into one of four classes

Slide by Ross Hardison, Penn State U.
De Bruin Graph

Nodes: k-mers
Edges: connects k-mers within **sequenced** k+1-mer on a short read

Slide by Sorin Istrail, Brown U.
Simplified De Bruin Graph
join all k-mers on a path without branching

Assembly = find Eulerian walk visiting each edge once
Edge-disjoint loops are a problem: multiple solutions

A graph can have multiple Eulerian walks, only one of which corresponds to the original superstring.

Right: graph for $ZABCDABEFABY$, $k=2$

Alternative Eulerian walks:

- $ZA \rightarrow AB \rightarrow BE \rightarrow EF \rightarrow FA \rightarrow AB \rightarrow BC \rightarrow CD \rightarrow DA \rightarrow AB \rightarrow BY$
- $ZA \rightarrow AB \rightarrow BC \rightarrow CD \rightarrow DA \rightarrow AB \rightarrow BE \rightarrow EF \rightarrow FA \rightarrow AB \rightarrow BY$

These correspond to two edge-disjoint directed cycles joined by node $AB$

$AB$ is a repeat: $ZABCDABEFABY$

Adapted from a slide by Ben Langmead, Johns Hopkins U.
How to assemble genome with repeats?

- **Answer:** longer reads
- **But:** cheap sequencing = short reads

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>700</td>
</tr>
<tr>
<td>Illumina</td>
<td>50–250</td>
</tr>
<tr>
<td>SOLiD</td>
<td>50</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>200</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>2900</td>
</tr>
<tr>
<td>Sanger</td>
<td>400–900</td>
</tr>
</tbody>
</table>

Example of disjoint loops
\[
q = 1 - p \quad \Rightarrow \quad p + q = 1 \quad \Rightarrow \quad (p + q)^n = 1
\]

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

\[
E(X) = \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x q^{n-x} = np
\]

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

Some trick:

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

\[
E(X^2) = E(X(X-1)) + E(X) = p^2 n(n-1) + np
\]

\[
(E[X])^2 = p^2 \frac{\partial^2}{\partial p^2} \rightarrow \text{Var}(X) = np - p^2 n = np(1-p)
\]
(a) Reads

1. ACCTGATC
2. CTGATCAA
3. TGATCAAT
4. AGCGATCA
5. CGATCAAT
6. GATCAATG
7. TCAATGTG
8. CAATGTGA

(b) Overlap graph

(c) de Bruijn graph of K-mers (5-mers shown here)

ACCTG ▶ CCTGA ▶ CTGAT ▶ TGATC ▶ GATCA ▶ ATCAA ▶ TCAAT ▶ CAATG ▶ AATGT ▶ ATGTG ▶ TGTGA
AGCGA ▶ GCGAT ▶ CGATC

(d) String graph
I showed my masterpiece to the grown-ups and asked them if my drawing frightened them.
Short interspersed repetitive elements: SINEs

• Example: Alu repeats
  – Most abundant repeated DNA in primates
  – Short, about 300 bp
  – About 1 million copies
  – Likely derived from the gene for 7SL RNA
  – Cause new mutations in humans

• They are retrotransposons
  – DNA segments that move via an RNA intermediate.

• MIRs: Mammalian interspersed repeats
  – SINES found in all mammals

• Analogous short retrotransposons found in genomes of all vertebrates.

Slide by Ross Hardison, Penn State U.
Long interspersed repetitive elements: LINEs

- Moderately abundant, long repeats
  - LINE1 family: most abundant
  - Up to 7000 bp long
  - About 50,000 copies
- Retrotransposons
  - Encode reverse transcriptase and other enzymes required for transposition
  - No long terminal repeats (LTRs)
- Cause new mutations in humans
- Homologous repeats found in all mammals and many other animals

Slide by Ross Hardison, Penn State U.
Other common interspersed repeated sequences in humans

• LTR-containing retrotransposons
  – MaLR: mammalian, LTR retrotransposons
  – Endogenous retroviruses
  – MER4 (MEedium Reiterated repeat, family 4)
• Repeats that resemble DNA transposons
  – MER1 and MER2
  – Mariner repeats
  – Were active early in mammalian evolution but are now inactive

Slide by Ross Hardison, Penn State U.
How many repeats are in eukaryotic genomes?

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 $C_0 l_{1/2}$ curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0 l_{1/2}$ 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
I showed my masterpiece to the grown-ups and asked them if my drawing frightened them.
Types of repeats are in humans

<table>
<thead>
<tr>
<th>Elements</th>
<th>Type</th>
<th>Number of elements*</th>
<th>Length occupied (bp)</th>
<th>Percentage of sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINEs</td>
<td></td>
<td>8</td>
<td>2093</td>
<td>4.19</td>
</tr>
<tr>
<td>ALUs</td>
<td></td>
<td>7</td>
<td>2011</td>
<td>4.02</td>
</tr>
<tr>
<td>MiRs</td>
<td></td>
<td>1</td>
<td>82</td>
<td>0.16</td>
</tr>
<tr>
<td>LINEs</td>
<td></td>
<td>16</td>
<td>12,279</td>
<td>24.56</td>
</tr>
<tr>
<td><strong>LINE1</strong></td>
<td></td>
<td><strong>12</strong></td>
<td><strong>11,419</strong></td>
<td><strong>22.84</strong></td>
</tr>
<tr>
<td>LINE2</td>
<td></td>
<td>4</td>
<td>860</td>
<td>1.72</td>
</tr>
<tr>
<td>L3/CR1</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LTR elements</td>
<td></td>
<td>5</td>
<td>1556</td>
<td>3.11</td>
</tr>
<tr>
<td>ERVL</td>
<td></td>
<td>1</td>
<td>513</td>
<td>1.03</td>
</tr>
<tr>
<td>ERVL-MaLRs</td>
<td></td>
<td>2</td>
<td>669</td>
<td>1.34</td>
</tr>
<tr>
<td>ERV_classI</td>
<td></td>
<td>2</td>
<td>374</td>
<td>0.75</td>
</tr>
<tr>
<td>ERV_classII</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DNA elements</td>
<td></td>
<td>1</td>
<td>248</td>
<td>0.5</td>
</tr>
<tr>
<td>hAT-Charlie</td>
<td></td>
<td>1</td>
<td>248</td>
<td>0.5</td>
</tr>
<tr>
<td>TcMar-Tigger</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Unclassified</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total interspersed repeats</td>
<td></td>
<td>Total</td>
<td>16,176</td>
<td>32.35</td>
</tr>
<tr>
<td>Small RNA</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Satellites</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>repeats</td>
<td>18</td>
<td>824</td>
<td>1.65</td>
</tr>
<tr>
<td>Low complexity</td>
<td></td>
<td>3</td>
<td>363</td>
<td>0.73</td>
</tr>
</tbody>
</table>

~ 1 kb each
Figure 1. Repetitive sequences in the human genome.

http://journals.plos.org/plosgenetics/article?id=info:doi/10.1371/journal.pgen.1003402
Summary: Parameters of a Probability Distribution

• The **mean**, \( \mu = E[X] \), is a measure of the **center of mass** of a random variable.

• The **variance**, \( V(X) = E[(X - \mu)^2] \), is a measure of the **dispersion** of a random variable around its mean.

• The **standard deviation**, \( \sigma = \sqrt{V(X)} \), is another measure of the dispersion around mean.

• The **skewness**, \( \gamma_1 = E[(X - \mu)^3/\sigma^3] \), is a measure of asymmetry around mean.

• The **geometric mean**, \( \exp(E[\log X]) \), is useful for very broad distributions.

• All can be infinite! Practically it means they increase with sample size.

• Different distributions can have identical parameters.