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 \((D/L)^*100\%\) 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>$P(X=x)$</th>
<th>0.6561</th>
<th>0.2916</th>
<th>0.0486</th>
<th>0.0036</th>
<th>0.0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$ = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$ = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$ = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$ = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X$ = 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$P(X \leq x)$</th>
<th>$P(X &gt; x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6561</td>
<td>0.3439</td>
</tr>
<tr>
<td>0.9477</td>
<td>0.0523</td>
</tr>
<tr>
<td>0.9963</td>
<td>0.0037</td>
</tr>
<tr>
<td><strong>0.9999</strong></td>
<td><strong>0.0001</strong></td>
</tr>
<tr>
<td><strong>1.0000</strong></td>
<td><strong>0.0000</strong></td>
</tr>
</tbody>
</table>

**Cumulative Distribution Function CDF:** $F(x) = P(X \leq x)$

Example:

$F(3) = P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = \textbf{0.9999}$

**Complementary Cumulative Distribution Function** (tail distribution) or **CCDF:** $F_>(x) = P(X > x)$

Example: $F_>(3) = P(X > 3) = 1 - P(X \leq 3) = \textbf{0.0001}$
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 integer and tail $P(X=x)>c/x^2$
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 \( f(x) \),

\[
E[h(X)] = \sum_x h(x) \cdot P(X = x) = \sum_x h(x) f(x) \tag{3-4}
\]

If \( h(x) = (X - \mu)^2 \), then its expectation, \( V(x) \), is the variance of \( X \).

\[
\sigma = \sqrt{V(x)}, \text{ is called standard deviation of } X
\]

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

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

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

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

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

Variance can be infinite if \( X \) can be any 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 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.

$$f(x_i) = \frac{1}{n}$$
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:

\[ f(x) = \frac{1}{b-a+1} \]

• Its measures are:

\[ \mu = E(x) = \frac{(b+a)}{2} \]

\[ \sigma^2 = V(x) = \frac{[(b-a+1)^2−1]}{12} \]

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?

- $C_t = 15 - 25$
- Random samples from test
- PMF($C_t$)
- Detectable at $C_t \leq 40$
- 5 days
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, https://academic.oup.com/ajcp/article/154/4/479/5873820
Why should we care? Variants are here

Figure 1: Simple frequency analysis of ORF gene Ct values from 641 positive samples, with a bin size of 0.5. Arrow and bar on the X-axis indicate nadir of frequency and Ct values, respectively. A Ct of 45 is displayed in the analysis and denotes “no signal detected” in the assay.

S-variant SARS-CoV-2 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-QPCR.

Michael Kidd1*, Alex Richter1, Angus Best1, Jeremy Mirza1, Benita Percival1, Megan Mayhew1, Oliver Megram1, Fiona Ashford1, Thomas White1, Emma Moles-Garcia1, Liam Crawford1, Andrew Bosworth1, Tim Plant1, Alan McNally1

1. Public Health England and University Hospitals Birmingham NHS Foundation Trust UK.