Review of probability theory and statistics

Xiaohui Xie
Department of Computer Science
University of California, irvine
Events and Sample space

Events as sets

- **Sample space** \((\Omega)\): the set of all possible outcomes of an experiment
- **Event** \((A)\): subset of the sample space, i.e., \(A \subseteq \Omega\)
- **Elementary event** \((\omega)\): member of \(\Omega\), i.e., \(\omega \in \Omega\)
Events and Sample space

Events as sets

- **Sample space** ($\Omega$): the set of all possible outcomes of an experiment
- **Event** ($A$): subset of the sample space, i.e., $A \subseteq \Omega$
- **Elementary event** ($\omega$): member of $\Omega$, i.e., $\omega \in \Omega$

Example: A coin is tossed repeatedly until the first head turns up.

- $\omega_i$: denotes the outcome where the first $i - 1$ tosses are tails and the $i$th toss is a head.
- Sample space: $\Omega = \{\omega_1, \omega_2, \cdots\}$
- An example event $A$: the first head occurs after an even number of tosses, i.e., $A = \{\omega_2, \omega_4, \cdots\}$
**σ-algebra**

**Definition**
A collection $\mathcal{F}$ of subsets of $\Omega$ is called a $\sigma$-algebra if it satisfies the following conditions:

- $\emptyset \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
- If $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
Definition
A collection \( \mathcal{F} \) of subsets of \( \Omega \) is called a \( \sigma \)-\textbf{algebra} if it satisfies the following conditions:

- \( \emptyset \in \mathcal{F} \)
- if \( A \in \mathcal{F} \) then \( A^c \in \mathcal{F} \)
- if \( A_1, A_2, \ldots \in \mathcal{F} \) then \( \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \)

Examples:
- \( \mathcal{F} = \{\emptyset, \Omega\} \): the smallest \( \sigma \)-field associated with \( \Omega \)
- \( \mathcal{F} = \{\emptyset, A, A^c, \Omega\} \) for any \( A \subset \Omega \)
- \( \{0, 1\}^\Omega \): the power set of \( \Omega \), containing all subsets of \( \Omega \)
**Probability measure**

**Definition**

A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ satisfying

- $\mathbb{P}(\Omega) = 1$;
- (countable additivity) if $A_1, A_2, \cdots$ is a collection of disjoint members of $\mathcal{F}$, then

\[
\mathbb{P}\left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).
\]

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.
Example: A coin, possibly biased, is tossed once.

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$
- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{H\}) = p$, $\mathbb{P}(\{T\}) = 1 - p$, $\mathbb{P}(\Omega) = 1$
Independence

Definition
Events $A$ and $B$ are called **independent** if

$$
\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)
$$

More generally, a family $\{A_i : i \in I\}$ is called **independent** if

$$
\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)
$$

for all finite subsets $J$ of $I$. 
Definition
If $\mathbb{P}(B) > 0$ then the **conditional probability** that $A$ occurs given that $B$ occurs is defined to be

$$
\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.
$$
Conditional Probability

Definition
If $\mathbb{P}(B) > 0$ then the **conditional probability** that $A$ occurs given that $B$ occurs is defined to be

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$ 

Let $B_1, B_2, \cdots, B_n$ be a partition of $\Omega$ such that $\mathbb{P}(B_i) > 0$ for all $i$. Then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$
Bayes’ theorem

If \( \mathbb{P}(B) > 0 \) then the probability of \( A \) given \( B \) is

\[
\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.
\]

Terminology:

- \( \mathbb{P}(A) \): the **prior probability** or **marginal probability** of \( A \).
- \( \mathbb{P}(B|A) \): the conditional probability of \( B \) given \( A \), also called the **likelihood**.
- \( \mathbb{P}(A|B) \): the conditional probability of \( A \) given \( B \), also called the **posterior probability**.
- \( \mathbb{P}(B) \): the **marginal probability** of \( B \), acting as a normalizing constant.
Example: false positives

A rare disease affects one person in $10^5$. A test for the disease shows positive with probability 0.99 when applied to an ill person, and with probability 0.01 when applied to a healthy person. What is the probability that you have the disease given that the test shows positive?

Solution:

$$\mathbb{P}(\text{ill} \mid +) = \frac{\mathbb{P}(+ \mid \text{ill}) \cdot \mathbb{P}(\text{ill})}{\mathbb{P}(+ \mid \text{ill}) \cdot \mathbb{P}(\text{ill}) + \mathbb{P}(+ \mid \text{healthy}) \cdot \mathbb{P}(\text{healthy})} = 0.99 \cdot \frac{1}{0.99 \cdot 10^{-5} + 0.01 \cdot 0.99999} \approx 0.001$$

Thus the false positive rate of the test will be very high.
Example: false positives

A rare disease affects one person in $10^5$. A test for the disease shows positive with probability 0.99 when applied to an ill person, and with probability 0.01 when applied to a healthy person. What is the probability that you have the disease given that the test shows positive?

Solution:

$$P(\text{ill}|+) = \frac{P(+|\text{ill})P(\text{ill})}{P(+|\text{ill})P(\text{ill}) + P(+|\text{healthy})P(\text{healthy})} \approx 0.001$$

Thus the false positive rate of the test will be very high.
Random variable

Definition
Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A random variable is a function \(X : \Omega \to \mathbb{R}\) with the property that for every Borel subset \(B\) of \(\mathbb{R}\), the subset of \(\Omega\)

\[
\{X \in B\} \equiv \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.
\]

\(X\) is said to be \(\mathcal{F}\)-measurable.

Without worrying about technical details, you can think of a random variable simply as a function mapping \(\Omega\) into \(\mathbb{R}\).
Random variable

Definition
Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A random variable is a function \(X : \Omega \to \mathbb{R}\) with the property that for every Borel subset \(B\) of \(\mathbb{R}\), the subset of \(\Omega\)

\[\{X \in B\} \equiv \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.\]

\(X\) is said to be \(\mathcal{F}\)-measurable.

Without worrying about technical details, you can think of a random variable simply as a function mapping \(\Omega\) into \(\mathbb{R}\).

Notations: Use upper-case letters (such as \(X, Y,\) and \(Z\)) to represent random variables, and lower-case letters (such as \(x, y,\) and \(z\)) to represent possible numerical values of these variables.
Random variable

Definition

The **distribution function** of a random variable $X$ is the function $F : \mathbb{R} \to [0, 1]$ given by

$$F(x) = \mathbb{P}(X \leq x),$$

where $\{X \leq x\}$ is an abbreviation of $\{\omega \in \Omega : X(\omega) \leq x\}$. 
Random variable

Definition
The **distribution function** of a random variable $X$ is the function $F : \mathbb{R} \rightarrow [0, 1]$ given by

$$F(x) = \mathbb{P}(X \leq x),$$

where \(\{X \leq x\}\) is an abbreviation of \(\{\omega \in \Omega : X(\omega) \leq x\}\).

Properties of a distribution function
- \(\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1;\)
- if \(x \leq y\) then \(F(x) \leq F(y);\)
- \(F\) is right-continuous, that is, \(\lim_{h \downarrow 0} F(x + h) = F(x).\)
Discrete and continuous variable

Definition
The random variable $X$ is called **discrete** if it takes values in some countable subset $\{x_1, x_2, \ldots, \}$, only, of $\mathbb{R}$. The discrete random variable $X$ has *(probability) mass function* $f : \mathbb{R} \to [0, 1]$ given by $f(x) = \mathbb{P}(X = x)$.

Definition
The random variable $X$ is called **continuous** if its distribution function can be expressed as

$$ F(x) = \int_{-\infty}^{x} f(u)du $$

for some integrable function $f : \mathbb{R} \to [0, \infty)$, called the *(probability) density function* of $X$. 
One Bernoulli trial

A coin is tossed once, and a head turns up (called a *success*) with probability $p$.

- $\Omega = \{H, T\}$
- Random variable $X$: Let $X$ to be 1 if the outcome is $H$, and 0 otherwise.
- Probability mass function of $X$:
  $$f(x) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$
- Mean: $\mathbb{E}(X) = p$
- Variance: $\text{var}(X) = p(1 - p)$

$X$ is said to have the **Bernoulli distribution** with parameters $p$, written as Bern($p$).
Binomial distribution

A coin is tossed $n$ times, and a head turns up each time with probability $p(=1-q)$.

$\Omega = \{H, T\}^n$

Random variable $X$: the total number of heads (successes)

Probability mass function of $X$:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n$$

Mean: $\mathbb{E}(X) = np$

Variance: $\text{var}(X) = np(1-p)$

$X$ is said to have the binomial distribution with parameters $n$ and $p$, written as $\text{bin}(n, p)$. 
Poisson distribution

If a random variable \( X \) takes values in the set \( \{0, 1, 2, \cdots\} \) with mass function

\[
f(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \cdots,
\]

where \( \lambda > 0 \), then \( X \) is said to have the **Poisson distribution** with parameter \( \lambda \).

- **Mean:** \( \mathbb{E}(X) = \lambda \)
- **Variance:** \( \text{var}(X) = \lambda \)
- The Poisson variable \( X \) can be viewed as a special binomial variable \( Y \sim \text{bin}(n, p) \) with \( n \to \infty \) and \( p \to 0 \), but with \( np = \lambda \) kept finite.
Hypergeometric distribution

Suppose that an urn contains $N$ objects, of which $m$ are white. Of these, $n$ objects are taken out of the urn at random without replacement. Let $X$ be the number of white objects taken out. Then $X$ is a random variable following the hypergeometric distribution.

- Probability mass function of $X$:

$$\mathbb{P}(X = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

- Support of $X$: $\max(0, n + m - N) \leq k \leq \min(m, n)$
- Mean: $\mathbb{E}(X) = \frac{nm}{N}$
- Variance: $\text{var}(X) = \frac{nm(N-n)(N-m)}{N^2(N-1)}$
Geometric distribution

Let $X$ be the number of Bernoulli trials (parameter $p$) needed to get one success (head), supported on the set $\{1, 2, \cdots\}$.

- Probability mass function of $X$: for some $p \in (0, 1)$
  \[ f(k) = p(1 - p)^{k-1}, \quad k = 1, 2, \cdots. \]

- Mean: $\mathbb{E}(X) = \frac{1}{p}$

- Variance: $\text{var}(X) = \frac{1-p}{p^2}$

The random variable $X$ is said to have the geometric distribution with parameter $p$. 
A random variable $X$ is **uniform** on $[a, b]$, written as $U(a, b)$, if it has probability density function

$$f(x) = \begin{cases} 
\frac{1}{b-a} & \text{if } a \leq x \leq b \\
0 & \text{otherwise.}
\end{cases}$$

- **Mean:** $\mathbb{E}(X) = \frac{b-a}{2}$
- **Variance:** $\text{var}(X) = \frac{(b-a)^2}{12}$
A random variable $X$ is **exponential** with parameter $\lambda (> 0)$, written as $X \sim \text{Exp}(\lambda)$, if it has probability density function

$$
f(x, \lambda) = \begin{cases} 
\lambda e^{-\lambda x} & \text{if } x \geq 0 \\
0 & \text{otherwise.}
\end{cases}
$$

- **Mean:** $\mathbb{E}(X) = \frac{1}{\lambda}$
- **Variance:** $\text{var}(X) = \frac{1}{\lambda^2}$
- $X$ describes the times between events in a Poisson process with rate $\lambda$
A random variable $X$ has the **normal distribution** with parameters $\mu$ and $\sigma^2$, written as $X \sim N(\mu, \sigma^2)$, if it has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

for $-\infty < x < \infty$.

- **Mean**: $\mathbb{E}(X) = \mu$
- **Variance**: $\text{var}(X) = \sigma^2$
- $X$ is said to have the **standard normal distribution** if $\mu = 0$ and $\sigma = 1$. 
A random variable $X$ has the **Cauchy distribution** if it has probability density function

$$f(x) = \frac{1}{\pi(1 + x^2)},$$

for $-\infty < x < \infty$.

- Mean: $\mathbb{E}(X) = 0$
- Variance: $\text{var}(X) = \infty$
Beta distribution

A random variable $X$ has the **Beta distribution** with parameters $\alpha$ and $\beta$, written as $X \sim \text{Beta}(\alpha, \beta)$, if it has probability density function

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} y^{\beta-1}, \quad 0 \leq x \leq 1.$$ 

where the beta function $B$ is

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

- **Mean:** $\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta}$
- **Variance:** $\text{var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$
Multinomial distribution

A random vector \( X = (X_1, X_2, \cdots, X_K) \) has the **multinomial distribution** with parameters \( n \) and \( p = (p_1, p_2, \cdots, p_K) \) satisfying \( p_i \geq 0 \) for all \( i \) and \( \sum_{i=1}^{K} p_i = 1 \), written as \( X \sim \text{Mult}(n, p) \), if it has probability density function

\[
f(x_1, \cdots, x_K; n, p) = \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}
\]

for non-negative integers \( x_1, \cdots, x_K \) satisfying \( \sum_{i=1}^{K} x_i = n \).

- **Marginal distribution** \( X_i \sim \text{Bin}(n, p_i) \).
- **Mean**: \( \mathbb{E}(X_i) = np_i \).
- **Covariance**: \( \text{Cov}(X_i, X_j) = -np_i p_j \).
Dirichlet distribution

A random vector \( X = (X_1, X_2, \cdots, X_K) \) has the **Dirichlet distribution** with parameters \( \alpha_1, \alpha_2, \cdots, \alpha_K > 0 \), written as \( X \sim \text{Dir}(\alpha) \), if it has probability density function

\[
f(x_1, \cdots, x_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i-1},
\]

for all \( x_1, \cdots, x_K \geq 0 \) satisfying \( x_1 + \cdots + x_K = 1 \), where

\[
B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}.
\]

Properties: Define \( \alpha_0 = \sum_{i=1}^{K} \alpha_i \).

- **Marginal distribution** \( X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i) \).
- **Mean**: \( \mathbb{E}(X_i) = \frac{\alpha_i}{\alpha_0} \).
- **Covariance**: \( \text{Cov}(X_i, X_j) = \frac{-\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)} \).
A random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ has the **multivariate normal distribution** with parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n} \succ 0$, written as $\mathbf{X} \sim N(\mu, \Sigma)$, if it has probability density function

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu) \Sigma^{-1} (\mathbf{x} - \mu) \right], \quad \mathbf{x} \in \mathbb{R}^n.$$ 

- **Mean:** $\mathbb{E}(\mathbf{X}) = \mu$
- **Covariance matrix:** $[\text{cov}(X_i, X_j)]_{i=1, \cdots, n; j=1, \cdots, n} = \Sigma$
Condition distribution of normal variables

Suppose the random vector $\mathbf{X} \in \mathbb{R}^n$ follows the multivariate norm distribution, $\mathbf{X} \sim N(\mu, \Sigma)$. Let $\mathbf{X} = [\mathbf{X}_1 \, \mathbf{X}_2]$, where $\mathbf{X}_1 \in \mathbb{R}^q$ and $\mathbf{X}_2 \in \mathbb{R}^{n-q}$. What is the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_1 = x_1$?
Condition distribution of normal variables

Suppose the random vector $\mathbf{X} \in \mathbb{R}^n$ follows the multivariate norm distribution, $\mathbf{X} \sim N(\mu, \Sigma)$. Let $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$, where $\mathbf{X}_1 \in \mathbb{R}^q$ and $\mathbf{X}_2 \in \mathbb{R}^{n-q}$. What is the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_1 = x_1$?

Answer: Let

$$
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},
$$

where $\mu_1 \in \mathbb{R}^q$ and $\Sigma_{11} \in \mathbb{R}^{q \times q}$. Then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = x_2$ is

$$
\mathbf{X}_1 | \mathbf{X}_2 = x_2 \sim N(\bar{\mu}, \bar{\Sigma}),
$$

where

$$
\bar{\mu} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \mu_2) \quad \text{and} \quad \bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21},
$$

where $\bar{\Sigma}$ is the Schur complement of $\Sigma_{22}$ in $\Sigma$. 
Consider a matrix $X \in S^n$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $A \in S^q$. If $|A| \neq 0$, the matrix $S = C - B^T A^{-1} B$ is called the **Schur complement** of $A$ in $X$.

The Schur complement comes in up in solving linear equations,

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1} BS^{-1} B^T A^{-1} & -A^{-1} BS^{-1} \\ -S^{-1} B^T A^{-1} & S^{-1} \end{bmatrix}.$$

That is, $S$ is the inverse of the 2, 2 block entry of the inverse of $X$.

- $|X| = |A| |S|$.
- $X \succ 0$ if and only if $A \succ 0$ and $S \succ 0$.
- If $A \succ 0$, then $X \succeq 0$ if and only if $S \succeq 0$. 
Condition distribution of normal variables

Let $\Omega = \Sigma^{-1}$, called *precision matrix*, be the inverse of the covariance matrix. Partition $\Omega$

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix},$$

where $\Omega_{11} \in \mathbb{R}^{q \times q}$. Then the conditional distribution of $X_1$ given $X_2 = x_2$ is

$$X_1 | X_2 = x_2 \sim N(\bar{\mu}, \bar{\Sigma}),$$

where

$$\bar{\mu} = \mu_1 - \Omega_{11}^{-1} \Omega_{12} (x_2 - \mu_2) \text{ and } \bar{\Sigma} = \Omega_{11}^{-1}.$$
Derived random variables

Suppose $X_1, \cdots, X_n$ are i.i.d. continuous random variables. Let

$$X_{\text{max}} = \max\{X_1, X_2, \cdots, X_n\}.$$ 

What is the distribution of $X_{\text{max}}$?
Suppose $X_1, \cdots, X_n$ are i.i.d. continuous random variables. Let

$$X_{\text{max}} = \max\{X_1, X_2, \cdots, X_n\}.$$ 

What is the distribution of $X_{\text{max}}$?

Let $f(x)$ and $F(x)$ be the probability density function and the cumulative distribution, respectively, of each $X_i$. Then

$$F_{\text{max}}(x) = P(X_{\text{max}} \leq x) = (F(x))^n.$$ 

So the density function of $X_{\text{max}}$ is

$$f_{\text{max}}(x) = nf(x)(F(x))^{n-1}.$$
Suppose $X_1, \cdots, X_n$ are i.i.d. continuous random variables with $X_i \sim \text{Exp}(\lambda)$ for all $i$. Then the cumulative distribution of $X_{\text{max}}$ is

$$F_{\text{max}}(x) = (1 - e^{-\lambda x})^n$$
Sequence-matching example

Suppose we have two small DNA sequences:

ggagactgtagacagctaatgctata
gaacgccctagccacgagcccttatc
* * * *** ** ***

Question: *Do the two sequences show significantly more similarity than what is expected from two arbitrary segments of DNA from two species?*
Statistical hypothesis testing

- A **statistical hypothesis testing** is a method of making statistical decisions using experimental data.
Statistical hypothesis testing

- A **statistical hypothesis testing** is a method of making statistical decisions using experimental data.
- A result is called **statistically significant** if it is unlikely to have occurred by chance.
A statistical hypothesis testing is a method of making statistical decisions using experimental data. A result is called statistically significant if it is unlikely to have occurred by chance. These decisions are almost always made using null-hypothesis tests, that is, ones that answer the question:

Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?
Hypothesis testing

General steps for hypothesis testing:

1. Declare null hypothesis ($H_0$) and alternative hypothesis ($H_1$). Let $p$ denote the probability of a match between the two nucleotides at any site.
Hypothesis testing

General steps for hypothesis testing:

1. Declare null hypothesis \((H_0)\) and alternative hypothesis \((H_1)\). Let \(p\) denote the probability of a match between the two nucleotides at any site.
   - \(H_0\): The two sequences were generated at random w.r.t. each other, i.e., \(p = 0.25\);
   - \(H_1\): \(p > 0.25\).
Hypothesis testing

General steps for hypothesis testing:

1. Declare null hypothesis (\(H_0\)) and alternative hypothesis (\(H_1\)). Let \(p\) denote the probability of a match between the two nucleotides at any site.
   - \(H_0\): The two sequences were generated at random w.r.t. each other, i.e., \(p = 0.25\);
   - \(H_1\): \(p > 0.25\).

2. Determine a test statistic.
   - Define \(Y\) to be the total number of matches.

Determine those observed values of the test statistic that lead to rejection of \(H_0\).

Obtain the data (all previous steps should be done before seeing the actual data), and determine whether the observed value belongs to those obtained in Step 3. If so, reject \(H_0\).
Hypothesis testing

General steps for hypothesis testing:

1. Declare null hypothesis ($H_0$) and alternative hypothesis ($H_1$). Let $p$ denote the probability of a match between the two nucleotides at any site.
   - $H_0$: The two sequences were generated at random w.r.t. each other, i.e., $p = 0.25$;
   - $H_1$: $p > 0.25$.

2. Determine a test statistic.
   - Define $Y$ to be the total number of matches.

3. Determine those observed values of the test statistic that lead to rejection of $H_0$. 
Hypothesis testing

General steps for hypothesis testing:

1. Declare null hypothesis ($H_0$) and alternative hypothesis ($H_1$). Let $p$ denote the probability of a match between the two nucleotides at any site.
   - $H_0$: The two sequences were generated at random w.r.t. each other, i.e., $p = 0.25$;
   - $H_1$: $p > 0.25$.

2. Determine a test statistic.
   - Define $Y$ to be the total number of matches.

3. Determine those observed values of the test statistic that lead to rejection of $H_0$.

4. Obtain the data (all previous steps should be done before seeing the actual data), and determine whether the observed value belongs to those obtained in Step 3. If so, reject $H_0$. 
In statistical hypothesis testing, there are two types of errors that can be drawn:

- Type I error ($\alpha$ error, false positive): reject $H_0$ when it is true.
- Type II error ($\beta$ error, false negative): accept $H_0$ when it is false.
Sensitivity vs. specificity

TP: true positive; TN: true negative; FP: false positive; FN: false negative

- true positive rate (sensitivity):

\[ TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \]

- false positive rate

\[ FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \]

- specificity (SPC, true negative rate)

\[ SPC = \frac{TN}{N} = 1 - FPR \]

- false discovery rate (FDR)

\[ FDR = \frac{FP}{FP + TP} \]
Hypothesis testing

General steps for hypothesis testing:
3. Determine those observed values of the test statistic that lead to rejection of $H_0$.

\[ \text{Choose a Type I error threshold, say 5%. Find a significance point } K \text{ such that } \text{Prob}(\text{Type I error}) = \text{Prob}(Y \geq K \mid p = 0.25) = 0.05. \]

\[ \text{Let } n = 26 \text{ be the length of the sequences } \text{Prob}(Y \geq K \mid p = 0.25) = \sum_{k=K}^{n} f(k), \text{where } f(k) \text{ is the prob. mass function of } \text{Bin}(n, 0.25). \text{From here, we have } K = 10. \]

\[ Y = 11. \text{ Thus we reject } H_0. \]
Hypothesis testing

General steps for hypothesis testing:

3. Determine those observed values of the test statistic that lead to rejection of \( H_0 \).
   - Choose a Type I error threshold, say 5%. Find a significance point \( K \) such that
     \[
     \text{Prob}(\text{Type I error}) = \text{Prob}(Y \geq K|p = 0.25) = 0.05
     \]

4. Obtain the data (all previous steps should be done before seeing the actual data), and determine whether the observed value belongs to those obtained in Step 3. If so, reject \( H_0 \).
   - \( Y = 11 \).
   Thus we reject \( H_0 \).
Hypothesis testing

General steps for hypothesis testing:

3. Determine those observed values of the test statistic that lead to rejection of $H_0$.
   - Choose a Type I error threshold, say 5%. Find a significance point $K$ such that
     \[
     \text{Prob}(\text{Type I error}) = \text{Prob}(Y \geq K|p = 0.25) = 0.05
     \]
   - Let $n = 26$ be the length of the sequences
     \[
     \text{Prob}(Y \geq K|p = 0.25) = \sum_{k=K}^{n} f(k),
     \]
     where $f(k)$ is the prob. mass function of $\text{Bin}(n, 0.25)$. From here, we have $K = 10$.
Hypothesis testing

General steps for hypothesis testing:

3. Determine those observed values of the test statistic that lead to rejection of $H_0$.
   - Choose a Type I error threshold, say 5%. Find a significance point $K$ such that
     \[
     \text{Prob}(\text{Type I error}) = \text{Prob}(Y \geq K | p = 0.25) = 0.05
     \]
   - Let $n = 26$ be the length of the sequences
     \[
     \text{Prob}(Y \geq K | p = 0.25) = \sum_{k=K}^{n} f(k),
     \]
     where $f(k)$ is the prob. mass function of $\text{Bin}(n, 0.25)$. From here, we have $K = 10$.

4. Obtain the data (all previous steps should be done before seeing the actual data), and determine whether the observed value belongs to those obtained in Step 3. If so, reject $H_0$.
   - $Y = 11$.

Thus we reject $H_0$. 
Hypothesis testing: *p*-values

Calculate the *p*-value in Step 4 for the observed value of the test statistic:

4. Obtain the data (*all previous steps should be done before seeing the actual data*), and determine the observed value *y* of the test statistic. Calculate *p*-value:

\[ p\text{-value} = \text{Prob}(Y \geq y | H_0) \]

- In the sequence-matching example, *Y* = 11.
- 
  \[ p\text{-value} = \sum_{k=11}^{n} f(k) = 0.04, \]

  where \( f(k) \) is the prob. mass function of bin(*n*, 0.25).

Because *p*-value < 0.05, we reject \( H_0 \).
A cautionary note

- A null hypothesis is **never proven** by such methods, as the absence of evidence against the null hypothesis does not establish its truth.

- In other words, one may either **reject**, or **not reject** the null hypothesis; one cannot **accept** it.

- This means that one cannot make decisions or draw conclusions that assume the truth of the null hypothesis.
Example 2: molecular classification of cancer

Expression of a gene called ATP6C (Vacuolar H+ ATPase proton channel subunit) in 38 bone marrow samples:

- 27 acute lymphoblastic leukemia (ALL):
  835 935 1665 764 1323 1030 1482 1306 593 2375 542 809
  2474 1514 1977 1235 933 1114 −9 3072 608 499 1740 1189
  1870 892 677

- 11 acute myeloid leukemia (AML):
  3237 1125 4655 3807 3593 2701 1737 3552 3255 4249 1870

Question: Is ATP6C differentially expressed in the two classes of samples?
A **t-test** is any statistical hypothesis test in which the test statistic follows a **Students t distribution** if the null hypothesis is true.
Students t-distribution

Suppose $X_1, X_2, \cdots, X_n$ are independent random variables that are normally distributed with mean $\mu$ and variance $\sigma^2$.

- Sample mean: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Sample variance: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$
- Define

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

$T$ has a Student’s t-distribution with $n - 1$ degrees of freedom (df)

The t-distribution looks like the standard normal distribution but with fatter tails, and becomes exact when $n \to \infty$. 
Independent one-sample t-test

Null hypothesis: the population mean is equal to a specified value $\mu_0$.

Test statistic:

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

where $\bar{X}_n$ is the sample mean and $S_n^2$ is the sample standard deviation. $df = n - 1$. 
Independent two-sample t-test: equal sample size, equal variance

Null hypothesis: the population mean is equal.
Test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} / \sqrt{2 / n}}$$

with

$$S_{X_1X_2} = \sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{2}}$$

where

- $S_{X_1}$ and $S_{X_2}$ are the sample standard deviation from each group
- $n$: the number of participants of each group
- Degree of freedom $df = 2n - 2$
Independent two-sample t-test: unequal sample size, equal variance

Null hypothesis: the population mean is equal.

Test statistic:

\[
T = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

with

\[
S_{X_1X_2} = \sqrt{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2} / n_1 + n_2 - 2
\]

where

- \(S_{X_1}\) and \(S_{X_2}\) are the sample standard deviation from each group
- \(n_1, n_2\): the number of participants of group 1 and group 2 respectively.
- Degree of freedom \(df = n_1 + n_2 - 2\)
Molecular classification of tumor example

Question: *Is gene ATP6C differentially expressed in AML vs ALL?*

**Null hypothesis:** the population mean is equal in two groups.

Perform independent two-sample t-test with unequal sample size:

- $n_1 = 27$ in ALL, $n_2 = 11$ in AML. Thus $df = 36$.
- The test statistic: $t = -6.2082$
- $p$-value: $p = 3.6660 \times 10^{-7}$

Therefore we reject the null hypothesis.