4.1 What is probability?

**Probability** is considered the measure of the likelihood that an event will occur. It is quantified as a number between 0 and 1, with 0 indicating impossibility and 1 certainty. The higher the probability of an event, the more likely that it will occur.

For example,

- When two dice are tossed simultaneously, the most likely outcome is 7, i.e. getting 7 is the event of highest probability.

- However, the most likely event does not always happen when two dice are tossed.

- Therefore, the accuracy of a probabilistic prediction increases as the number of the experiments is increased. Theoretically speaking, when the experiment is repeated infinitely many times, the frequency of such an occurrence is expected to match exactly the probability.
Two major interpretations:

(1) **Objective interpretation:** Probabilities are numbers assigned to describe some objective or physical state affairs. *Frequentist probability* denotes the relative frequency of occurrence of an experiment’s outcome, when repeated a large number of times. *Propensity probability* is interpreted as the tendency of an experiment to yield a certain outcome, even if it is performed only once.

(2) **Subjective interpretation:** Probabilities are numbers assigned to express one’s opinion regarding how certain event is to occur. Subjective opinion has been linked to the “price at which one would buy or sell a bet that pays 1 unit of utility if E, 0 if not E.”
4.2 Useful terminology of set

Events frequently encountered in probabilistic theory are not always numbers. Concepts related to sets are useful since the collection of all possible outcomes of a probabilistic experiment is typically defined as a set.

<table>
<thead>
<tr>
<th>Set language</th>
<th>Set notation</th>
<th>Venn diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>The set</td>
<td>$\Omega$</td>
<td></td>
</tr>
<tr>
<td>Subset $A$</td>
<td>$A$ ( $A \subset \Omega$ )</td>
<td></td>
</tr>
<tr>
<td>Complement of $A$</td>
<td>$A^c$ ( $\Omega \setminus A$ )</td>
<td></td>
</tr>
<tr>
<td>Union of $A$, $B$</td>
<td>$A \cup B$</td>
<td></td>
</tr>
<tr>
<td>Intersection of $A$, $B$</td>
<td>$A \cap B$</td>
<td></td>
</tr>
</tbody>
</table>

When two subsets $A$, $B$ are disjoint (mutually exclusive), then $A \cap B = \emptyset$, where $\emptyset$ represents an empty set.
E.g. Tossing a die. All possible outcomes are represented by a discrete set of variables.

\[ \Omega = \{1, 2, 3, 4, 5, 6\} \]

The subset of all outcomes that are larger than 2 and that contains only even numbers are

\[ A = \{3, 4, 5, 6\}, \quad B = \{2, 4, 6\} \]

Therefore, the respective complements are

\[ A^c = \{1, 2\}, \quad B^c = \{1, 3, 5\} \]

And that,

\[ A \cup B = \{2, 3, 4, 5, 6\}, \quad A \cap B = \{4, 6\} \]

For a fair die, the probability of each possible outcome is equally \(1/6\). Thus,

\[
\begin{align*}
P(A) &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}, \\
P(A \cup B) &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}, \\
P(A \cap B) &= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}.
\end{align*}
\]
4.3 Basic rules in probability theory

1. Independence rule: If $A$ and $B$ are independent events, then

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B).$$

E.g. The probability of tossing a coin an then a die and yield head and 6 is

$$P(H \text{ and } 6) = P(H)P(6) = \frac{11}{26} = \frac{1}{12}.$$  

This is because the two events are independent of each other.

2. Complement rule:

$$P(\text{not } A) = P(A^c) = 1 - P(A).$$

E.g. The probability of tossing a coin an then a die and yield anything else but the head and 6 is

$$P(\text{not } H \text{ and } 6) = 1 - P(H \text{ and } 6) = 1 - \frac{1}{12} = \frac{11}{12}.$$
3. **Mutually none-exclusive rule:** If $A$ and $B$ are not mutually exclusive, then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

This is easy to understand since the common subset $A \cap B$ is added twice when adding $P(A)$ to $P(B)$. As a special case when $A$ and $B$ are mutually exclusive, i.e. $A \cap B = \emptyset$, then we have the mutually exclusive rule

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B).$$

4. **Difference rule:** If $A$ is a subset of $B$ (i.e. $A \subset B$), then

$$P(B \text{ but not } A) = P(B \cap A^c).$$

5. **Conditional probability and multiplication rule:** Let $P(A|B)$ and $P(B|A)$ be, respectively, the probability of $A$ occurs given that $B$ has occurred and vice versa. These are referred to as conditional probabilities. Then,

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B).$$

Therefore, conditional probabilities can be expressed as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
6. **Bayes’ rule:** The two conditional probabilities $P(A|B)$ and $P(B|A)$ are related by

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

**E.g.** Probability of having green eyes is $P(G) = 0.1$, of having brown hairs is $P(B) = 0.75$, and of having green eyes and brown hairs is $P(G \cap B) = 0.09$. Then, probability of having

(i) green eyes but not brown hairs is

$$P(G \cap B^c) = P(G) - P(G \cap B) = 0.1 - 0.09 = 0.01.$$

(ii) green eyes and/or brown hairs is

$$P(G \cup B) = P(G) + P(B) - P(G \cap B) = 0.1 + 0.75 - 0.09 = 0.76.$$

(iii) brown hairs given that a person has green eyes is

$$P(B|G) = \frac{P(B \cap G)}{P(G)} = \frac{0.09}{0.1} = 0.9.$$

(iv) green eyes given that a person has brown hairs is

$$P(G|B) = \frac{P(B|G)P(G)}{P(B)} = \frac{0.9 \times 0.1}{0.75} = 0.12.$$
### Summary of probabilities

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$P(A) \in [0, 1]$</td>
</tr>
<tr>
<td>not A</td>
<td>$P(A^C) = 1 - P(A)$</td>
</tr>
</tbody>
</table>
| A or B      | $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  
              | $P(A \cup B) = P(A) + P(B)$   if A and B are mutually exclusive |
| A and B     | $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$  
              | $P(A \cap B) = P(A)P(B)$      if A and B are independent |
| A given B   | $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$ |
4.4 Discrete random variables and probability distributions

A discrete random variable \((X)\) is a function that maps a sample domain \((\Omega)\) onto a countable set of numbers

\[
X : \Omega \rightarrow S \quad (S \subset \mathbb{R}),
\]

where \(X = \{X_1, X_2, \cdots, X_N\}\) takes a set of \(N\) values.

**E.g. 4.4.1** Consider the experiment of tossing two fairs dice.

The domain contains all possible outcomes of such an experiment:

\[
\Omega = \{(1,1), \cdots, (1,6), \cdots, (6,1), \cdots, (6,6)\} \quad (36 \text{ in total})
\]

Let the random variable \(X\) be defined as a map \(X : \Omega \rightarrow \mathbb{R}\) such that

\[
X(i, j) = i + j, \quad \text{(the sum of the two numbers yielded)}
\]

where \(i, j = 1, 2, 3, 4, 5, 6\) and \(X\) takes on \(N = 11\) distinct values \(X = \{2, 3, \cdots, 12\}\). Therefore,

\[
P(X = 3) = P((1, 2) \text{ or } (2, 1)) = \frac{1}{36} + \frac{1}{36} = \frac{1}{18},
\]

\[
P(X \leq 3) = P((1, 2) \text{ or } (2, 1) \text{ or } (1, 1)) = \frac{3}{36} = \frac{1}{12}.
\]
Discrete probability distribution: Suppose \( X = \{X_1, X_2, \cdots, X_N\} \), the probability function associated with \( X \) is

\[
P_i = P(X = X_i), \quad (i = 1, 2, \cdots, N, \ 0 \leq P_i \leq 1)
\]

is called a discrete probability distribution. It must satisfy

\[
\sum_{i=1}^{N} P_i = P_1 + P_2 + \cdots + P_N = 1.
\]

E.g. 4.4.1 Continued! The probability distribution for the experiment of tossing two fair dice can be obtained as follows.

<table>
<thead>
<tr>
<th>i</th>
<th>Samples/Events</th>
<th>( X_i = m + n )</th>
<th>( P_i = P(X_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{(1,1)}</td>
<td>2</td>
<td>1/36</td>
</tr>
<tr>
<td>2</td>
<td>{(1,2),(2,1)}</td>
<td>3</td>
<td>2/36</td>
</tr>
<tr>
<td>3</td>
<td>{(1,3),(3,1),(2,2)}</td>
<td>4</td>
<td>3/36</td>
</tr>
<tr>
<td>4</td>
<td>{(1,4),(4,1),(2,3),(3,2)}</td>
<td>5</td>
<td>4/36</td>
</tr>
<tr>
<td>5</td>
<td>{(1,5),(5,1),(2,4),(4,2),(3,3)}</td>
<td>6</td>
<td>5/36</td>
</tr>
<tr>
<td>6</td>
<td>{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)}</td>
<td>7</td>
<td>6/36</td>
</tr>
<tr>
<td>7</td>
<td>{(2,6),(6,2),(3,5),(5,3),(4,4)}</td>
<td>8</td>
<td>5/36</td>
</tr>
<tr>
<td>8</td>
<td>{(3,6),(6,3),(4,5),(5,4)}</td>
<td>9</td>
<td>4/36</td>
</tr>
<tr>
<td>9</td>
<td>{(4,6),(6,4),(5,5)}</td>
<td>10</td>
<td>3/36</td>
</tr>
<tr>
<td>10</td>
<td>{(5,2),(6,5)}</td>
<td>11</td>
<td>2/36</td>
</tr>
<tr>
<td>11</td>
<td>{(6,6)}</td>
<td>12</td>
<td>1/36</td>
</tr>
</tbody>
</table>
The graph of the probability distribution:

Expected value: Let $X = \{X_1, X_2, \cdots, X_N\}$ be a discrete random variable with probability distribution $P = \{P_1, P_2, \cdots, P_N\}$. Then, its expected value is defined as:

$$
\mu = \overline{X} = E(X) = \sum_{i=1}^{N} X_i P_i.
$$

- Intuitively, the expected value is the long-run average of repetitions of the experiment it represents.
- The law of large numbers states that the arithmetic mean of the values almost surely converges to the expected value as the number of repetitions approaches infinity.
• Expected value of a constant is equal to the constant itself: if \( C = \text{constant} \), then \( \overline{C} = E(C) = \sum_{i=1}^{N} CP_i = C \sum_{i=1}^{N} P_i = C \).

• Calculating expected value is a linear operation: let \( X, Y \) be two sets of random variables with identical probability distribution. Then,

\[
E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y), \quad (\alpha, \beta \in \mathbb{R}).
\]

• The expected value often is not in the set of all possible values of the random variable. E.g. the random variable representing all possible outcomes of tossing a fair die can take the following values: \( X = \{1, 2, 3, 4, 5, 6\} \). However, \( E(X) = 3.5 \) which is not one of those values.

---

Two important discrete probability distributions:

**I.** Bernoulli distribution \( B(1, p) \) for Bernoulli trial: Simply speaking, Bernoulli trial is an experiment with a probability \( p \) of yielding 1 (success) and \( 1 - p \) of yielding 0 (failure).

To make the trial more physically understandable, one can consider a Brownian particle placed at the origin on the axis of real numbers. A coin is tossed with a probability \( p \) of getting a head and \( 1 - p \) of getting a tail. The particle is moved 1 unit to the right when yielding a head and remains still when getting a tail. (Anybody can actually do this experiment with a coin and a piece of rock.)

• B – Bernoulli;

• 1 – experiment done once/coin toss once;
• p – probability of success or yielding H or moving the particle 1 unit to the right.

In this experiment, \( \Omega = \{H, T\} \) is the sample space, \( X^{(1,p)} = \{X_1(H) = 1, X_2(T) = 0\} \) is the random variable that can be regarded as the distance between the particle and the origin.

The probability distribution is:
\[
P = \{P_1, P_2\} = \{P_1(H) = p, P_2(T) = 1 - p\} = \{p, 1 - p\}.
\]

Thus, the expected value is
\[
E(X^{(1,p)}) = 1 \cdot p + 0 \cdot (1 - p) = p.
\]

**Variance** of a random variable is defined as
\[
\text{Var}(X) = E[(X - E(X))^2] = E[X^2 - 2\mu X + \mu^2]
= E[X^2] - 2\mu E[X] + \mu^2 = E(X^2) - \mu^2.
\]

Therefore, the variance of the random variable \( X^{(1,p)} \) is
\[
\text{Var}(X^{(1,p)}) = E[(X^{(1,p)})^2] - [E(X^{(1,p)})]^2
= 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p - p^2 = p(1 - p).
\]

• Notice that the variance of the sum of two random variables is equal to the sum of the variances of the two random variables:
\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).
\]
II. Bernoulli distribution $B(n, p)$ for $n$ independent Bernoulli trials:

- B — Bernoulli;
- n — experiment done $n$ times /coin toss $n$ times;
- p — probability of success or yielding H or moving the particle 1 unit to the right.

The sample space:

$$\Omega = \{(H^0, T^n), (H^1, T^{n-1}), \ldots, (H^n, T^0)\}, \quad (N = n + 1).$$

The random variable is defined as

$$X^{(n,p)} = \text{number of success after } n \text{ trials} = \text{distance moved after } n \text{ coin toss.}$$

Thus,

$$X^{(n,p)} = \{0, 1, \ldots, n\}, \quad (N = n + 1).$$

The probability distribution is

$$P_k = P(X^{(n,p)} = k) = P(H^k, T^{n-k}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad (k = 0, 1, \ldots, n)$$

where

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

is the total number of ways to pick $k$ boxes out of a total of $n$ boxes. It is often called the binomial coefficient. The distribution is also called the binomial distribution. This result is based on the following observations.
- Each possible outcome giving rise to \( k \) heads (e.g. getting \( k \) heads in the first \( k \) tosses and \( n - k \) tails in the following tosses) has a probability of \( p^k(1 - p)^{n-k} \).

- But there is a total of \( \binom{n}{k} \) outcomes giving rise to \( k \) heads each with a probability of \( p^k(1 - p)^{n-k} \).

**The Binomial Theorem:** Let \( a \), \( b \) be two real numbers. Then,

\[
(a + b)^n = (a + b)(a + b) \cdots (a + b) = \sum \text{product of } n \text{ numbers each picked from separate brackets} = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = \sum_{k=0}^{n} \binom{n}{k} b^k a^{n-k}.
\]

To make this formula easier to understand, we outline a few specific cases of the theorem.

\[
(a + b)^0 = 1 \quad (a + b)^1 = a + b \quad (a + b)^2 = a^2 + 2ab + b^2 \\
\vdots \\
(a + b)^n = \binom{n}{0} a^n + \binom{n}{1} a^{n-1}b + \cdots + \binom{n}{n-1} ab^{n-1} + \binom{n}{n} b^n
\]
Notice that $B(n, p)$ is equivalent to $B(1, p)$ repeated independently $n$ times. Thus,

- $X^{(n,p)} = X_1^{(1,p)} + X_2^{(1,p)} + \cdots + X_n^{(1,p)}$.

- As a consequence,

\[
E(X^{(n,p)}) = E(X_1^{(1,p)} + X_2^{(1,p)} + \cdots + X_n^{(1,p)}) \\
= E(X_1^{(1,p)}) + \cdots + E(X_n^{(1,p)}) \\
= p + \cdots + p = np.
\]

\[
\text{Var}(X^{(n,p)}) = \text{Var}(X_1^{(1,p)} + \text{Var}_2^{(1,p)} + \cdots + \text{Var}_n^{(1,p)}) \\
= \text{Var}(X_1^{(1,p)}) + \cdots + \text{Var}(X_n^{(1,p)}) \\
= p(1 - p) + \cdots + p(1 - p) = np(1 - p).
\]
4.5 Genetic drift and Wright-Fisher model

**Genetic (Allelic) drift:** is the change in the frequency of an existing gene variant (allele) in a population due to random sampling of organism.

- Alleles in the offsprings samples of the alleles in the parent generation.
- Allele frequency is defined as the fraction of the copies of one gene that share a common structure/form.
- Genetic drift may cause gene variants to disappear and thereby reduce genetic variation.

**Wright-Fisher model:** basic assumptions

- A population of (a large number) \( N \) haploid individuals.
- No generation overlap.
- Population size \( N \) remains fixed from generation from generation.
- A gene with alleles \( A \) and \( a \) with no selection (i.e. no difference in fitness between the two).

**Random variables:**

\[ X_t = \text{number of individuals with A allele at time } t. \]
\[ p_t = \frac{X_t}{N} = \text{frequency of A allele at time } t. \]
Probability distribution is given by

\[ P(X_{t+1} = k) = \binom{N}{k} p_t^k (1 - p_t)^{N-k}, \quad (0 \leq k \leq N). \]

- This is binomial distribution.
- This is the probability of picking \( k \) individuals with \( A \) allele and \( n - k \) individuals with \( a \) allele to reproduce.