Many things in life are uncertain. Can we ‘measure’ and compare such uncertainty so that it helps us to make more informed decision? Probability theory provides a systematic way of doing so.

1. **Basic Set Theory**

The basic language of probability theory is provided by a branch of mathematics called set theory. Even though it has a lot of fascinating stories in it, we will only be needing the most basic concepts.

A set is a collection of abstract elements. If a set \( \Omega \) contains \( n \) elements \( x_1, x_2, \cdots, x_n \), then we write

\[
\Omega = \{ x_1, x_2, \cdots, x_n \}.
\]

For each \( i = 1, 2, \cdots, n \), we write \( x_i \in \Omega \), meaning that \( x_i \) is an element of \( \Omega \). A subcollection \( A \) of the elements of \( \Omega \) is called a subset of \( \Omega \), and we write \( A \subseteq \Omega \) or \( \Omega \supseteq A \). If \( x_i \) is an element of \( A \) we write \( x_i \in A \); otherwise, \( x_i \notin A \). When we describe \( A \), either we list all of its elements as in (1), or we use the following conditional statement

\[
A = \{ x \in \Omega \mid x \text{ has property } P \}.
\]

For instance, if \( \Omega = \{1,2,3,4,5,6\} \) and \( A = \{2,4,6\} \), then we can also write

\[
A = \{ x \in \Omega \mid x \text{ is even} \}.
\]

\( \Omega \) is a subset of itself. A subset of \( \Omega \) containing no element is called the empty set, and is denoted by \( \emptyset \).

Let \( A, B \) be two subsets of \( \Omega \). Define their union \( A \cup B \) and intersection \( A \cap B \) by

\[
A \cup B = \{ x \in \Omega \mid x \in A \text{ or } x \in B \} \quad (4)
\]
\[
A \cap B = \{ x \in \Omega \mid x \in A \text{ and } x \in B \}. \quad (5)
\]

We write \( A = B \) if they consists of the same elements.

**Exercise 1.1.** Let \( \Omega \) be a set and let \( A, B \subseteq \Omega \). Show that \( A = B \) if and only if \( A \subseteq B \) and \( B \subseteq A \).

**Exercise 1.2.** Let \( \Omega \) be a set and let \( A \subseteq \Omega \). Define the **complement** of \( A \), denoted by \( A^c \), by

\[
A^c = \{ x \in \Omega \mid x \notin A \} \quad (6)
\]

Show that \( A \cup A^c = \Omega \) and \( A \cap A^c = \emptyset \).

**Exercise 1.3.** Let \( \Omega \) be a set and let \( A, B \subseteq \Omega \). Define a subset \( A \setminus B \) of \( \Omega \) by

\[
A \setminus B = \{ x \in \Omega \mid x \in A \text{ and } x \notin B \}. \quad (7)
\]

Show that \( A \setminus B = A \cap B^c \).

Union and intersection can be defined among more than two subsets of \( \Omega \). Let \( A_1, A_2, \cdots, A_k \) be subsets of \( \Omega \). Define the union and intersection of \( A_i \)'s by

\[
\bigcup_{i=1}^{k} A_i = \{ x \in \Omega \mid x \in A_i \text{ for some } i \in \{1,2,\cdots,k\} \}, \quad (8)
\]
\[
\bigcap_{i=1}^{k} A_i = \{ x \in \Omega \mid x \in A_i \text{ for all } i \in \{1,2,\cdots,k\} \}. \quad (9)
\]
Exercise 1.4 (de Morgan’s law). Let \( \Omega \) be a set and let \( A_1, \cdots, A_k \subseteq \Omega \). Show that
\[
\left( \bigcup_{i=1}^{k} A_i \right) = \left( \bigcap_{i=1}^{k} A_i^c \right).
\] (10)

Exercise 1.5. Let \( \Omega \) be a set and let \( B, A_1, A_2 \cdots \subseteq \Omega \). Show that
\[
B \cap \left( \bigcup_{i=1}^{\infty} A_i \right) = \bigcap_{i=1}^{\infty} B \cap A_i,
\] (11)
\[
B \cup \left( \bigcap_{i=1}^{\infty} A_i \right) = \bigcap_{i=1}^{\infty} B \cup A_i.
\] (12)

Given two sets \( A \) and \( B \), we can form their cartesian product \( A \times B \) by the set of all pairs of elements in \( A \) and \( B \). That is,
\[
A \times B = \{(a, b) \mid a \in A, b \in B\}.
\] (13)

We write \( A^2 = A \times A \), \( A^3 = A \times A \times A \), and so on.

Lastly, all of the above discussion can be extended when the grounding set \( \Omega \) consists of infinitely many elements. For instance, the set of all integers \( \mathbb{Z} \), the set of all natural numbers \( \mathbb{N} \), and the set of all real numbers \( \mathbb{R} \). A set \( \Omega \) containing infinitely many elements is said to be countably infinite if there is a one-to-one correspondence between \( \Omega \) and \( \mathbb{N} \); it is said to be uncountably infinite otherwise. One of the most well-known example of uncountably infinite set is \( \mathbb{R} \), which is proved by the celebrated Cantor’s diagonalization argument.

Exercise 1.6 (Binary expansion). Let \([0,1] \subseteq \mathbb{R}\) be the unit interval, which is also called the continuum.

(i) Given an element \( x \in [0,1] \), show that there exists a unique binary expansion of \( x \). That is, there exists a sequence of integers \( x_1, x_2, \cdots \) from \([0,1]\) such that
\[
x = 0.x_1x_2x_3\cdots
\] (14)
\[
:= x_1 + \frac{x_2}{2} + \frac{x_3}{2^3} + \cdots.
\] (15)
(Hint: Divide \([0,1]\) into \([0,1/2) \cup [1/2,1]\). Then \( x_1 = 0 \) if \( x \) belongs to the first half, and \( x_1 = 1 \) otherwise. Subdivide the interval and see which half it belongs to, and so on.)

(ii) Given a binary expansion \( 0.x_1x_2x_3\cdots \), define a sequence of real numbers \( y_1, y_2, \cdots \) by
\[
y_n = 0.x_1x_2\cdots x_n
\] (16)
\[
= \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n}.
\] (17)
Show that the sequence \((y_n)_{n \geq 1}\) is non-decreasing and bounded above by 1. Conclude that there exists a limit \( y := \lim_{n \to \infty} y_n \).

Remark 1.7. Binary expansion in fact gives a one-two-one correspondence between the unit interval \([0,1]\) and the infinite product \([0,1]^{\mathbb{N}} = [0,1] \times [0,1] \times \cdots \) of 0’s and 1’s.

Exercise 1.8 (Cantor’s diagonalization argument). Let \([0,1] \subseteq \mathbb{R}\) be the unit interval.

(i) Suppose \([0,1]\) is countably infinite. Then we can enumerate all of its elements by \( a_1, a_2, a_3, \cdots \).

Using (i), we can write each \( a_i \)’s by its unique binary expansion:
\[
a_1 = 0.a_{11}a_{12}a_{13}a_{14}\cdots
\] (18)
\[
a_2 = 0.a_{21}a_{22}a_{23}a_{24}\cdots
\] (19)
\[
a_3 = 0.a_{31}a_{32}a_{33}a_{34}\cdots
\] (20)
\[
a_4 = 0.a_{41}a_{42}a_{43}a_{44}\cdots
\] (21)
Conclude that \([0, 1]\) is uncountably infinite. Since \([0,1]\)

\[
\alpha = 0.a_1 a_2 a_3 a_4 \cdots ,
\]

where \(\hat{0} = 1\) and \(\hat{1} = 0\). Show that \(\alpha \neq a_i\) for all \(i \geq 1\). Hence we have found an element of \([0,1]\) that is not among the list \(a_1, a_2, a_3, \cdots\).

(ii) Conclude that \([0, 1]\) is uncountably infinite. Since \([0, 1] \subseteq \mathbb{R}\), it follows that \(\mathbb{R}\) is also uncountably infinite.

2. Probability measure and probability space

We begin with idealizing our situation. Let \(\Omega\) be a finite set, called sample space. This is the collection of all possible outcomes that we can observe (think of six sides of a die). We are going to perform some experiment on \(\Omega\), and the outcome could be any subset \(E\) of \(\Omega\), which we call an event. Let us denote the collection of all events \(E \subseteq \Omega\) by \(2^{\Omega}\). A probability measure on \(\Omega\) is a function \(\mathbb{P} : 2^{\Omega} \rightarrow [0, 1]\) such that for each event \(E \subseteq \Omega\), it assigns a number \(\mathbb{P}(E) \in [0, 1]\) and satisfies the following properties:

(i) \(\mathbb{P}(\emptyset) = 0\) and \(\mathbb{P}(\Omega) = 1\).

(ii) If two events \(E_1, E_2 \subseteq \Omega\) are disjoint, then \(\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_1)\).

Remark 2.1. For general sample space \(\Omega\) (not necessarily finite), not every subset of \(\Omega\) can be an event. Precise definition of the collection of ‘events’ for the general case is beyond the scope of this course. On the other hand, the axiom (ii) for the probability measure needs to be replaced with the following countable version:

(ii)’ For a countable collection of disjoint events \(A_1, A_2, \cdots \subseteq \Omega\),

\[
\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{24}
\]

Exercise 2.2. Let \(\mathbb{P}\) be a probability measure on sample space \(\Omega\). Show the following.

(i) Let \(E = \{x_1, x_2, \cdots, x_k\} \subseteq \Omega\) be an event. Then \(\mathbb{P}(E) = \sum_{i=1}^{k} \mathbb{P}(\{x_i\}) = 1\).

(ii) \(\sum_{x \in \Omega} \mathbb{P}(\{x\}) = 1\).

If \(\mathbb{P}\) is a probability measure on sample space \(\Omega\), we call the pair \((\Omega, \mathbb{P})\) a probability space. This is our idealized world where we can precisely measure uncertainty of all possible events. Of course, there could be many (in fact, infinitely many) different probability measures on the same sample space.

Exercise 2.3 (coin flip). Let \(\Omega = \{H, T\}\) be a sample space. Fix a parameter \(p \in [0, 1]\), and define a function \(\mathbb{P}_p : 2^{\Omega} \rightarrow [0, 1]\) by \(\mathbb{P}_p(\emptyset) = 0, \mathbb{P}_p(\{H\}) = p, \mathbb{P}_p(\{T\}) = 1 - p, \mathbb{P}_p(\{H, T\}) = 1\). Verify that \(\mathbb{P}_p\) is a probability measure on \(\Omega\) for each value of \(p\).

A typical way of constructing a probability measure is to specify how likely it is to see each individual element in \(\Omega\). Namely, let \(f : \Omega \rightarrow [0, 1]\) be a function that sums up to 1, i.e., \(\sum_{x \in \Omega} f(x) = 1\). Define a function \(\mathbb{P} : 2^{\Omega} \rightarrow [0, 1]\) by

\[
\mathbb{P}(E) = \sum_{\omega \in E} f(\omega). \tag{25}
\]

Then this is a probability measure on \(\Omega\), and \(f\) is called a probability distribution on \(\Omega\). For instance, the probability distribution on \(\{H, T\}\) we used to define \(\mathbb{P}_p\) in Exercise 2.3 is \(f(H) = p\) and \(f(T) = 1 - p\).
Example 2.4 (Uniform probability measure). Let $\Omega = \{1, 2, \cdots, m\}$ be a sample space and let $P$ be the uniform probability measure on $\Omega$, that is,

$$P(\{x\}) = 1/m \quad \forall x \in \Omega.$$  \hfill (26)

Then for the event $A = \{1, 2, 3\}$, we have

$$P(A) = P(\{1\} \cup \{2\} \cup \{3\})$$

$$= P(\{1\}) + P(\{2\}) + P(\{3\})$$

$$= \frac{1}{m} + \frac{1}{m} + \frac{1}{m} = \frac{3}{m}$$  \hfill (27)

Likewise, if $A \subseteq \Omega$ is any event and if we let $|A|$ denote the size (number of elements) of $A$, then

$$P(A) = \frac{|A|}{m}.$$  \hfill (30)

For example, let $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ be the sample space of a roll of two fair dice. Let $A$ be the event that the sum of two dice is 5. Then

$$A = \{(1,4), (2,3), (3,2), (4,1)\},$$


Exercise 2.5. Show that the function $P : 2^\Omega \to [0,1]$ defined in (25) is a probability measure on $\Omega$. Conversely, show that every probability measure on a finite sample space $\Omega$ can be defined in this way.

Remark 2.6 (General probability space). A probability space does not need to be finite, but we need a more careful definition in that case. For example, if we take $\Omega$ to be the unit interval $[0, 1]$, then we have to be careful in deciding which subset $E \subseteq \Omega$ can be an 'event': not every subset of $\Omega$ can be an event. A proper definition of general probability space is out of the scope of this course.

Exercise 2.7. Let $(\Omega, P)$ be a probability space and let $A \subseteq \Omega$ be an event. Show that $P(A^c) = 1 - P(A)$.

Example 2.8 (Roll of two dice). Suppose we roll two dice and let $X$ and $Y$ be the outcome of each die. Say all possible joint outcomes are equally likely. The sample space for the roll of a single die can be written as $\{1, 2, 3, 4, 5, 6\}$, so the sample space for rolling two dice can be written as $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. In picture, think of 6 by 6 square grid and each node represents a unique outcome $(x, y)$ of the roll. By assumption, each out come has probability 1/36. Namely,

$$P((X, Y) = (x, y)) = 1/36 \quad \forall 1 \leq x, y \leq 6.$$  \hfill (32)

This gives the uniform probability distribution on our sample space $\Omega$ (see Example 2.4).

We can compute various probabilities for this experiment. For example,

$$P(\text{at least one die is even}) = 1 - P(\text{both dice are odd})$$

$$= 1 - P(\{1, 3, 5\} \times \{1, 3, 5\}) = 1 - \frac{9}{36} = \frac{3}{4},$$  \hfill (33)

where for the first equality we have used the complementary probability in Exercise 2.7.

Now think about the following question: *What is the most probable value for the sum $X + Y$?* By considering diagonal lines $x + y = k$ in the 2-dimensional plane for different values of $k$, we find

$$P(X + Y = k) = \frac{\# \text{ of intersections between the line } x + y = k \text{ and } \Omega}{36}.$$  \hfill (35)

From example, $P(X + Y = 2) = 1/36$ and $P(X + Y = 7) = 6/36 = 1/6$. Moreover, from Figure 1, it is clear that the number of intersections is maximized when the diagonal line $x + y = k$ passes
Figure 1. Sample space representation for roll of two independent fair dice and and events of fixed sum of two dice.

through the extreme points (1,6) and (6,1). Hence 7 is the most probable value for $X + Y$ with the probability being 1/6.

Exercise 2.9 (Roll of three dice). Suppose we roll three dice and all possible joint outcomes are equally likely. Identify the sample space $\Omega = \{1,2,3,4,5,6\}^3$ as a (6 x 6 x 6) 3-dimensional integer lattice, and let $X, Y,$ and $Z$ denote the outcome of each die.

(i) Write down the probability distribution on $\Omega$.

(ii) For each $k \geq 1$, show that

$$P(X + Y + Z = k) = \frac{\text{# of intersections between the plane } x + y + z = k \text{ and } \Omega}{6^3}.$$  

What are the minimum and maximum possible values for $X + Y + Z$?

(iii) Draw a cube for $\Omega$ and planes $x + y + z = k$ for $k = 3, 5, 10, 11, 16, 18$. Argue that the intersection gets larger as $k$ increases from 3 to 10 and smaller as $k$ goes from 11 to 18. Conclude that 10 and 11 are the most probable values for $X + Y + Z$.

(iv) Consider the following identity

$$(x + x^2 + x^3 + x^4 + x^5 + x^6)^3 = x^{18} + 3x^{17} + 6x^{16} + 10x^{15} + 15x^{14} + 21x^{13} + 25x^{12} + 27x^{11} + 27x^{10} + 25x^9 + 21x^8 + 15x^7 + 10x^6 + 6x^5 + 3x^4 + x^3$$

Show that the coefficient of $x^k$ in the right hand side equals the size of the intersection between $\Omega$ and the plane $x + y + z = k$. Conclude that

$$P(X + Y + Z = 10) = P(X + Y + Z = 11) = \frac{27}{6^3} = \frac{1}{8}.$$  

(This way of calculating probabilities is called the generating function method.)

Exercise 2.10. Suppose Nate commutes to campus by Bruin bus, which arrives at his nearby bus stop every 10 min. Suppose each bus waits at the stop for 1 min. What is the probability that Nate takes no more than 3 min at the stop until he takes a bus? (Hint: Represent the sample space as a unit square in the coordinate plane)

The following are important properties of probability measure.

Theorem 2.11. Let $(\Omega, \mathbb{P})$ be a probability space. These followings hold:
(i) (Monotonicity) For any events $A \subseteq B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(ii) (Subadditivity) For $A \subseteq \bigcup_{i=1}^{\infty} A_i$, $\mathbb{P}(A) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

(iii) (Continuity from below) If $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$, then $\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n)$.

(iv) (Continuity from above) If $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \bigcap_{i=1}^{\infty} A_i$, then $\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n)$.

Proof. (i) Since $A \subseteq B$, write $B = A \cup (B \setminus A)$. Note that $A$ and $B \setminus A$ are disjoint. Hence by the second axiom of probability measure, we get

$$
\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A),
$$

where the last inequality uses the fact that $\mathbb{P}(B \setminus A) \geq 0$.

(ii) The events $A_i$'s are not necessarily disjoint, but we can cook up a collection of disjoint events $B_i$'s so that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Namely, we define

$$
B_1 = A_1 \subseteq A_1
$$

$$
B_2 = (A_1 \cup A_2) \setminus A_1 \subseteq A_2
$$

$$
B_3 = (A_1 \cup A_2 \cup A_3) \setminus (A_1 \cup A_2) \subseteq A_3,
$$

and so on. Then clearly $B_i$'s are disjoint and their union is the same as the union of $A_i$'s. Now by part (i) and axiom (ii)' of probability measure (See Remark 2.1), we get

$$
\mathbb{P}(A) \leq \mathbb{P} \left( \bigcup_{i=1}^{\infty} A_i \right) = \mathbb{P} \left( \bigcup_{i=1}^{\infty} B_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).
$$

(iii) Define a collection of disjoint subsets $B_i$'s similarly as in (ii). In this case, they will be

$$
B_1 = A_1 \subseteq A_1
$$

$$
B_2 = A_2 \setminus A_1 \subseteq A_2
$$

$$
B_3 = A_3 \setminus A_2 \subseteq A_3
$$

and so on. Then $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ and $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$. Hence we get

$$
\mathbb{P}(A) = \mathbb{P} \left( \bigcup_{i=1}^{\infty} A_i \right) = \mathbb{P} \left( \bigcup_{i=1}^{\infty} B_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i)
$$

$$
= \lim_{k \to \infty} \sum_{i=1}^{k} \mathbb{P}(B_i) = \lim_{n \to \infty} \mathbb{P}(B_1 \cup \cdots \cup B_n) = \lim_{n \to \infty} \mathbb{P}(A_n).
$$

(iv) Note that $A_1^c \subseteq A_2^c \subseteq \cdots$ and $A^c = \bigcup_{i=1}^{\infty} A_i^c$ by de Morgan's law (Exercise 1.4). Hence by part (iii), we deduce

$$
\mathbb{P}(A^c) = \lim_{n \to \infty} \mathbb{P}(A_n^c).
$$

But then by Exercise 2.7, this yields

$$
\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \lim_{n \to \infty} \mathbb{P}(A_n^c) = \lim_{n \to \infty} (1 - \mathbb{P}(A_n^c)) = \lim_{n \to \infty} \mathbb{P}(A_n).
$$

Some immediate consequences of Theorem 2.11 (ii) are given in the following exercise.

**Exercise 2.12** (Union bound). Let $(\Omega, \mathbb{P})$ be a probability space.

(i) For any $A, B \subseteq \Omega$ such that $A \subseteq B$, show that

$$
\mathbb{P}(A) \leq \mathbb{P}(B).
$$

(ii) For any $A, B \subseteq \Omega$, show that

$$
\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).
$$
Then the conditional probability $P$ and this quantity is called the Exercise 2.13 

(iii) Let $A_1, A_2, \ldots, A_k \subseteq \Omega$. By an induction on $k$, show that

$$P \left( \bigcup_{i=1}^{k} A_i \right) \leq \sum_{i=1}^{k} P(A_i).$$ (55)

(iv) (Countable subadditivity) Let $A_1, A_2, \ldots \subseteq \Omega$ be a countable collection of events. Show that

$$P \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} P(A_i).$$ (56)

**Exercise 2.13** (Inclusion-exclusion). Let $(\Omega, P)$ be a probability space. Let $A_1, A_2, \ldots, A_k \subseteq \Omega$. Show the following.

(i) For any $A, B \subseteq \Omega$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$ (57)

(ii) For any $A, B, C \subseteq \Omega$,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$ (58)

(iii)* Let $A_1, A_2, \ldots, A_k \subseteq \Omega$. Use an induction on $k$ to show that

$$P \left( \bigcup_{i=1}^{k} A_i \right) = \sum_{i=1}^{k} P(A_{i_1}) - \sum_{1 \leq i_1 < i_2 \leq k} P(A_{i_1} \cap A_{i_2})$$

$$+ \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq k} P(A_{i_1} \cap \cdots \cap A_{i_k}) - \cdots + (-1)^k P \left( \bigcap_{i=1}^{k} A_i \right).$$ (60)

**Remark 2.14.** Later we will show the general inclusion-exclusion in a much easier way using random variables and expectation.

### 3. Conditional probability

Consider two experiments on a probability space and the outcomes are recorded by $X$ and $Y$. For instance, $X$ could be the number of friends on Facebook and $Y$ could be the number of connections on LinkedIn of a randomly chosen classmate. Perhaps it would be case that $Y$ is large if $X$ is large. Or maybe the opposite is true. In any case, the outcome of $Y$ is most likely be affected by knowing something about $X$. This leads to the notion of ‘conditioning’. For any two events $E_1$ and $E_2$ such that $P(E_2) > 0$, we define

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$ (61)

and this quantity is called the **conditional probability** of $E_1$ given $E_2$. Note that $P(E_1 \mid E_2) \leq 1$ since $E_1 \cap E_2 \subseteq E_2$ and

**Example 3.1.** Roll a fair die and let $X \in \{1, 2, 3, 4, 5, 6\}$ be the outcome. Then

$$P(X = 2) = 1/6,$$ (62)

$$P(X = 2 \mid X \text{ is even}) = \frac{P(X = 2)}{P(X \in \{2, 4, 6\})} = 1/3,$$ (63)

$$P(X = 2 \mid X \in \{3, 4\}) = \frac{P(\emptyset)}{P(\{3, 4\})} = 0.$$ (64)

\[ \blacktriangle \]

In the following statement, we show that conditioning on a fixed event induces a valid probability measure.

**Proposition 3.2.** Let $(\Omega, P)$ be a probability space, and let $B \subseteq \Omega$ be an event such that $P(B) > 0$. Then the conditional probability $P(\cdot \mid B)$ is indeed a probability measure on $\Omega$. 


Proof. We need to verify the two axioms of probability measure. First, note that

\[ P(\emptyset | B) = \frac{P(\emptyset \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0, \]

(65)

\[ P(\Omega | B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1. \]

(66)

This verifies axiom (i). For axiom (ii), let \( A_1, A_2, \ldots \) be disjoint events. Then

\[ P \left( \bigcup_{i=1}^{\infty} A_i | B \right) = \frac{P \left( \bigcup_{i=1}^{\infty} (A_i \cap B) \right)}{P(B)} \]

(67)

\[ = \frac{P \left( \bigcup_{i=1}^{\infty} A_i \cap B \right)}{P(B)} \]

(68)

\[ = \frac{1}{P(B)} \sum_{i=1}^{\infty} P(A_i \cap B) \]

(69)

\[ = \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \]

(70)

\[ = \sum_{i=1}^{\infty} P(A_i | B). \]

(71)

Note that the second equality uses Exercise 1.5; The third one uses the fact that \( A_i \cap B \)'s are disjoint (since \( A_i \)'s are) and axiom (ii) for the original probability measure \( P \). The last equality uses definition of conditional probability.

Example 3.3. Consider rolling two four-sided dice and let \((X, Y)\) be the outcome in the sample space \( \Omega = \{0, 1, 2, 3\}^2 \). Suppose the two dice somehow affect each other according to the probability distribution on \( \Omega \) is depicted in Figure 2.

![Figure 2](image.png)

**Figure 2.** Joint distribution on \( \Omega \) shown in red. Common denominator of 33 is omitted in the figure.

Then we can compute the conditional probability \( P(X \geq 2 | Y = 2) \) as below.

\[ P(X \geq 2 | Y = 2) = \frac{P(X = 2 | Y = 2) + P(X = 3 | Y = 2)}{P(Y = 2)} \]

(72)

\[ = \frac{\frac{5/33}{P(Y = 2)} + \frac{1/33}{P(Y = 2)}}{(3 + 0 + 5 + 1)/33} = \frac{2/3}{(3 + 0 + 5 + 1)/33} = 2/3. \]

(74)
Example 3.4. Consider tossing a fair coin three times. Identify the sample space as $\Omega = \{H, T\}^3$, where we suppose that all possible outcomes are equally likely with probability 1/8. Consider the following two events:

$$A = \{\text{more heads than tails come up}\}, \quad B = \{\text{second toss is a tail}\}. \quad (75)$$

Then note that

$$A = \{\text{two heads}\} \cup \{\text{three heads}\} \quad (76)$$

$$= \{(H, H, T), (H, T, H), (T, H, H)\} \cup \{(H, H, H)\}, \quad (77)$$

so by Exercise 2.4,

$$\mathbb{P}(A) = \frac{4}{8} = 1/2. \quad (78)$$

Now suppose we know that the event $B$ occurs. How does this knowledge affects the probability of $A$? The new probability of $A$ is given by the conditional probability $\mathbb{P}(A|B)$. To compute this, note that

$$B = \{(T, H, T), (T, H, H), (H, H, T), (H, H, H)\}, \quad A \cap B = \{(T, H, H), (H, H, T), (H, H, H)\}. \quad (79)$$

Hence

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{3/8}{4/8} = 3/4. \quad (80)$$

Furthermore, we can determine the full conditional probability measure $\mathbb{P}(\cdot | B)$. Recall that conditioning on the event $B$ changes the sample space $\Omega$ to $B$. In order to describe a probability measure, we only need to specify its distribution (see Exercise 2.5). In this case, the distribution for $\mathbb{P}(\cdot | B)$ is given by

$$\mathbb{P}((T, H, T)|B) = \frac{\mathbb{P}((T, H, T))}{\mathbb{P}(B)} = 1/4, \quad (81)$$

$$\mathbb{P}((T, H, H)|B) = \frac{\mathbb{P}((T, H, H))}{\mathbb{P}(B)} = 1/4, \quad (82)$$

$$\mathbb{P}((H, H, T)|B) = \frac{\mathbb{P}((H, H, T))}{\mathbb{P}(B)} = 1/4, \quad (83)$$

$$\mathbb{P}((H, H, H)|B) = \frac{\mathbb{P}((H, H, H))}{\mathbb{P}(B)} = 1/4. \quad (84)$$

Hence $\mathbb{P}(\cdot | B)$ is the uniform probability measure on $B$. \hfill \blacktriangle

4. Partitioning the sample space and Bayes’ theorem

4.1. Partitioning the sample space. Sometimes when we try to compute the probability of a certain event $A$, a ‘divide and conquer’ approach could be very useful. Namely, we divide the sample space into smaller and disjoint events $A_i$, and compute the probability of $A$ conditioned on $A_i$, and then we add up the results. This technique of computing probability is called partitioning.

Proposition 4.1 (Partitioning). Let $(\Omega, \mathbb{P})$ be a probability space and let $A_1, A_2, \cdots, A_k \subseteq \Omega$ be a partition of $\Omega$. That is, $A_i$’s are disjoint and $\bigcup_{i=1}^k A_i = \Omega$. Then for any event $A \subseteq \Omega$,

$$\mathbb{P}(A) = \mathbb{P}(A|A_1)\mathbb{P}(A_1) + \mathbb{P}(A|A_2)\mathbb{P}(A_2) + \cdots + \mathbb{P}(A|A_k)\mathbb{P}(A_k). \quad (85)$$

Proof. Since $\bigcup_{i=1}^k A_i = \Omega$, we have

$$A = A \cap \Omega = A \cap \left( \bigcup_{i=1}^k A_i \right) = \bigcup_{i=1}^k A \cap A_i. \quad (86)$$
By definition of conditional probability,

\[ P(A \cap A_i) = P(A | A_i) P(A_i) \quad (87) \]

for each \( 1 \leq i \leq k \). Moreover, \( A \cap A_i \)'s are disjoint since \( A_i \)'s are. Hence

\[ P(A) = P\left( \bigcup_{i=1}^{k} A \cap A_i \right) = \sum_{i=1}^{k} P(A \cap A_i) = \sum_{i=1}^{k} P(A | A_i) P(A_i). \quad (88) \]

Here are some examples.

**Example 4.2.** Suppose you role a fair die repeatedly until the first time you see an odd number. What is the probability that the total sum of outcomes is exactly 5?

Before we proceed, let us first identify the sample space \( \Omega \) and probability measure \( P \). Note that we are not sure how many times we will have to roll the die. Here are some outcomes \( \omega \in \Omega \) of the experiment to get started:

\[
\begin{align*}
\text{(1)} \\
(2, 4, 5) \\
(2, 6, 6, 2, 4, 3).
\end{align*}
\]

Hence \( \Omega \) consists of the sequence of numbers from \( \{1, 2, 3, 4, 5, 6\} \) which consists of even numbers for all but the last coordinate. On the other hand, Observe that

\[
\begin{align*}
P((\{1\})) &= \frac{1}{6}, \\
P((\{2, 4, 5\})) &= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{6^3}, \\
P((\{2, 6, 6, 2, 4, 3\})) &= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{6^4}.
\end{align*}
\]

Hence, we observe that

\[ P(\omega) = \frac{1}{6^{|\omega|}}, \quad (95) \]

where \( |\omega| \) denotes the number of coordinates in \( \omega \).

In order to use partitioning, we first need to decide how we partition the entire sample space \( \Omega \). Let us partition the sample space according to the first outcome; let \( A_i \) be the event that first roll gives number \( i \). Then \( P(A_i) = 1/6 \) for all \( 1 \leq i \leq 6 \). On the other hand,

\[
\begin{align*}
P(A | A_1) &= 0, \\
P(A | A_3) &= 0, \\
P(A | A_5) &= 1 \\
\end{align*}
\]

since the roll ends after the first step if \( i \) is odd. For the other cases, note that

\[
\begin{align*}
P(A | A_2) &= P((\{2, 2, 1\} | A_2)) = \frac{1}{6^2} \\
P(A | A_4) &= P((\{4, 1\} | A_4)) = \frac{1}{6} \\
P(A | A_6) &= 0.
\end{align*}
\]

Thus we have

\[ P(A) = \sum_{i=1}^{6} P(A | A_i) P(A_i) \quad (102) \]

\[ = 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} = \frac{6^2 + 6 + 1}{6^3} = \frac{43}{6^3}. \quad (103) \]
Exercise 4.3 (Laplace’s rule of succession). Laplace ‘computed’ the probability that the sun will rise tomorrow, given that it has risen for the preceding 5000 years. The combinatorial model is as follow. Suppose we have \( N \) different coins, where the \( k \)th coin has probability \( k/N \) of coming up heads. We choose one of the \( N \) coin uniformly at random and flip it \( n \) times. For each \( 1 \leq i \leq n \), let \( R_i \) be the event that the \( i \)th flip comes up heads. We are interested in the following conditional probability

\[
P(R_{n+1} \mid R_n \cap R_{n-1} \cap \cdots \cap R_1). \tag{104}
\]

If we think of coin coming up heads as the event of sun rising, then this is a model for the probability that the sun will rise tomorrow given that it has risen for the past \( n \) days.

(i) Write \( R'_n = R_n \cap R_{n-1} \cap \cdots \cap R_1 \). Show that

\[
P(R_{n+1} \mid R'_n) = \frac{P(R_{n+1} \cap R'_n)}{P(R'_n)} = \frac{P(R'_{n+1})}{P(R'_n)}, \tag{105}
\]

(ii) For each \( 1 \leq i \leq n + 1 \), use partitioning to show that

\[
P(R'_i) = \sum_{k=1}^N P(R'_i \mid \text{prob. of heads is } k/N)P(\text{prob. of heads is } k/N) = \sum_{k=1}^N \left( \frac{k}{N} \right)^i \frac{1}{N}. \tag{106}
\]

(iii) By considering the upper and lower Riemann sums we have

\[
\sum_{k=1}^N \left( \frac{k}{N} \right)^i \frac{1}{N} \leq \int_0^1 t^i \, dt \leq \sum_{k=0}^{N-1} \left( \frac{k}{N} \right)^i \frac{1}{N}. \tag{107}
\]

Using (ii) and (iii), show that

\[
P(R'_i) \approx \int_0^1 t^i \, dt = \frac{1}{i+1}, \tag{108}
\]

where the approximation becomes exact as \( N \to \infty \).

(iv) From (i), conclude that

\[
P(R_{n+1} \mid R'_n) \approx \frac{n + 1}{n + 2}. \tag{109}
\]

For \( n = 5000 \times 365 \) (days), we have \( P(R_{5001} \mid R'_{5000}) \approx 0.9999994520 \). So it’s pretty likely that the sun will rise tomorrow as well.

4.2. Inference and Bayes’ Theorem. If we have a model for bitcoin price, then we can tune the parameters accordingly and attempt to predict its future prices. But how do we tune the parameters? Say the bitcoin price suddenly drops by 20% overnight (which is no surprise). Which factor is the most likely to have caused this drop? In general, Bayes’ Theorem can be used to infer likely factors when we are given the effect or outcome. This is all based on conditional probability and partitioning.

We begin by the following trivial but important observation:

\[
P(B \mid A)P(A) = P(A \cap B) = P(A \mid B)P(B). \tag{110}
\]

And this is all we need.

Theorem 4.4 (Bayes’ Theorem). Let \((\Omega, \mathbb{P})\) be a probability space, and let \( A_1, \cdots, A_k \subseteq \Omega \) be events of positive probability that form a partition of \( \Omega \). Then

\[
P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B)} = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)}. \tag{111}
\]

Proof. Note that (110) yields the first equality. Then partitioning and rewrite \( P(B) \) gives the second equality. \( \square \)
What’s so important about the Bayes’ theorem is its interpretation as a means of inference, which is one of the fundamental tools in modern machine learning.

**Example 4.5.** Suppose Bob has a coin with unknown probability of heads, which we denote by $\Theta$. This is called the parameter. Suppose Alice knows that Bob has one of the three kinds of coins $A$, $B$, and $C$, with probability of heads being $p_A = 0.2$, $p_B = 0.5$, and $p_C = 0.8$, respectively. This piece of information is called the model. Since Alice has no information, she initially assumes that Bob has one of the three coins equally likely. Namely, she assumes the uniform distribution over the sample space $\Omega = \{0.2, 0.5, 0.8\}$. This knowledge is called prior.

Now Bob flips his coin 10 times and got 7 heads, and reports this information, which we call Data, to Alice. Now that Alice has more information, she needs to update her prior to posterior using the new data. This will be likely to give higher weight to $\Theta = 0.2$.

Next, Alice needs to compute yet another posterior distribution to analyze the new data. This is called the posterior parameter $\Theta$. Hence we can compute the posterior distribution by

$$
\Pr(\Theta | \text{Data}) = \frac{\Pr(\text{Data} | \Theta) \Pr(\Theta)}{\Pr(\text{Data})}.
$$

This can be done by partitioning:

$$
\Pr(\text{Data}) = \Pr(\text{Data} | \Theta = 0.2) \Pr(\Theta = 0.2) + \Pr(\text{Data} | \Theta = 0.5) \Pr(\Theta = 0.5) + \Pr(\text{Data} | \Theta = 0.8) \Pr(\Theta = 0.8)
$$

$$
= \Pr(7/10 \text{ heads} | \Theta = 0.2) \frac{1}{3} + \Pr(7/10 \text{ heads} | \Theta = 0.5) \frac{1}{3} + \Pr(7/10 \text{ heads} | \Theta = 0.8) \frac{1}{3}
$$

$$
= \frac{1}{3} \left( \binom{10}{7} (0.2)^7 (0.8)^3 + \binom{10}{7} (0.5)^7 (0.5)^3 + \binom{10}{7} (0.8)^7 (0.2)^3 \right) \approx 0.1064,
$$

where $\binom{10}{7} = 120$ is the number of ways to choose 7 out of 10 objects.

Second, we reformulate the first equality of Theorem 4.4 as

$$
\frac{\Pr(\text{Data} | \Theta) \Pr(\Theta)}{\Pr(\text{Data})} = \frac{\Pr(\text{Data} | \Theta = 0.2) \Pr(\Theta = 0.2)}{\Pr(\text{Data})} = \frac{\Pr(\text{Data} | \Theta = 0.5) \Pr(\Theta = 0.5)}{\Pr(\text{Data})} = \frac{\Pr(\text{Data} | \Theta = 0.8) \Pr(\Theta = 0.8)}{\Pr(\text{Data})}.
$$

Hence we can compute the posterior distribution by

$$
\Pr(\Theta = 0.2 | \text{Data}) = \frac{\Pr(\text{Data} | \Theta = 0.2) \Pr(\Theta = 0.2)}{\Pr(\text{Data})} = \frac{\binom{10}{7} (0.2)^7 (0.8)^3}{0.1064} \approx 0.0025
$$

$$
\Pr(\Theta = 0.5 | \text{Data}) = \frac{\Pr(\text{Data} | \Theta = 0.5) \Pr(\Theta = 0.5)}{\Pr(\text{Data})} = \frac{\binom{10}{7} (0.5)^7 (0.5)^3}{0.1064} \approx 0.3670
$$

$$
\Pr(\Theta = 0.8 | \text{Data}) = \frac{\Pr(\text{Data} | \Theta = 0.8) \Pr(\Theta = 0.8)}{\Pr(\text{Data})} = \frac{\binom{10}{7} (0.8)^7 (0.2)^3}{0.1064} \approx 0.6305.
$$

Note that according to the posterior distribution, $\Theta = 0.8$ is the most likely value, which is natural given that we have 7 heads out of 10 flips. However, our knowledge is always incomplete so our posterior knowledge is still a probability distribution on the sample space.

What if Bob flips his coin another 10 times and reports only 3 heads to Alice? Then she will have to use her current prior $\pi = [0.0025, 0.3060, 0.6305]$ (which was obtained as the posterior in the previous round) to compute yet another posterior using the new data. This will be likely to give higher weight to $\Theta = 0.2$.

**Exercise 4.6.** Suppose we have a prior distribution $\pi = [0.0025, 0.3680, 0.6305]$ on the sample space $\Omega = \{0.2, 0.5, 0.8\}$ for the inference problem of unknown parameter $\Theta$. Suppose we are given the data that ten independent flips of probability $\Theta$ coin comes up heads twice. Compute the posterior distribution using this data and Bayesian inference.

**Exercise 4.7.** A test for pancreatic cancer is assumed to be correct 95% of the time: if a person has the cancer, the test results in positive with probability 0.95, and if the person does not have the cancer, then the test results in negative with probability 0.95. From a recent medical research, it is
known that only \%0.05 of the population have pancreatic cancer. Given that the person just tested positive, what is the probability of having the cancer?

5. INDEPENDENCE

When knowing something about one event does not yield any information of the other, we say the two events are independent. To make this statement a bit more precise, suppose Bob is betting $5 for whether an event $E_1$ occurs or not. Suppose Alice tells him that some other event $E_2$ holds true. If Bob can somehow leverage on this knowledge to increase his chance of winning, then we should say the two events are not independent.

Formally, we say two events $E_1$ and $E_2$ are independent if

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2). \quad (120)$$

We say two events or RVs are dependent if they are not independent.

**Example 5.1.** Flip two fair coins at the same time and let $X$ and $Y$ be their outcome, labeled by $H$ and $T$. Clearly knowing about one coin does not give any information of the other. For instance, the first coin lands on heads with probability $1/2$. Whether the first coin lands on heads or not, the second coin will land on heads with probability $1/2$. So

$$\mathbb{P}(X = H \text{ and } Y = H) = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(X = H)\mathbb{P}(Y = H). \quad (121)$$

$$\Box$$

**Exercise 5.2.** Suppose two events $A_1$ and $A_2$ are independent and suppose that $\mathbb{P}(A_2) > 0$. Show that

$$\mathbb{P}(A_1 | A_2) = \mathbb{P}(A_1). \quad (122)$$

In terms of the inference problem, the above equation can be written as

$$\mathbb{P}(\Theta | \text{Data}) = \mathbb{P}(\Theta). \quad (123)$$

That is, after observing Data, our prior distribution does not change. This is because the new data does not contain any new information.

**Example 5.3.** Consider rolling two 4-sided dice, where all 16 outcomes are equally likely. Let $\Omega = \{1,2,3,4\}^2$ be the sample space. Let $X$ and $Y$ be the outcome of the two dice.

(i) Note that for any $(i, j) \in \Omega$,

$$\mathbb{P}(X = i \text{ and } Y = j) = \mathbb{P}((i, j)) = \frac{1}{16} \quad (124)$$

$$\mathbb{P}(X = i) = \mathbb{P}((i, 1), (i, 2), (i, 3), (i, 4)) = \frac{1}{4} \quad (125)$$

$$\mathbb{P}(X = j) = \mathbb{P}((1, j), (2, j), (3, j), (4, j)) = \frac{1}{4}. \quad (126)$$

Hence $\mathbb{P}(X = i \text{ and } Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$ and the events $\{X = i\}$ and $\{Y = j\}$ are independent.

(ii) Let $A = \{X = 1\}$ and $B = \{X + Y = 5\}$. Are these events independent? It seems like knowing that $A$ occurs yields some information regarding $B$. However, it turns out that conditioning on $A$ does not change the probability of $B$ occurring. Namely,

$$\mathbb{P}(A \cap B) = \mathbb{P}((1, 4)) = \frac{1}{16} \quad (127)$$

$$\mathbb{P}(A) = \mathbb{P}((1, 1), (1, 2), (1, 3), (1, 4)) = \frac{1}{4} \quad (128)$$
\[ P(B) = P((1, 4), (2, 3), (3, 2), (4, 1)) = \frac{1}{4}. \]  

Hence \( P(A \cap B) = P(A)P(B) \), so they are independent. In other words,

\[
P(B|A) = \frac{P((1, 4))}{P(((1, 1), (1, 2), (1, 3), (1, 4)))} = \frac{1}{4} = P(((1, 4), (2, 3), (3, 2), (4, 1))) = P(B).
\]

(iii) Let \( A = \{\min(X, Y) = 2\} \) and \( B = \{\max(X, Y) = 2\} \). Then

\[ P(A \cap B) = P((2, 2)) = \frac{1}{16} \]  

\[ P(A) = P((2, 2), (2, 4), (3, 2), (4, 2)) = \frac{5}{16} \]  

\[ P(B) = P((1, 2), (2, 2), (2, 1)) = \frac{3}{16}, \]

so \( P(A \cap B) \neq P(A)P(B) \). Hence they are not independent.

\[ \blacklozenge \]

There is a conditional version of the notion of independence. Let \( A, B, C \) be events such that \( C \) has positive probability. We say \( A \) and \( B \) are independent conditional on \( C \) if

\[ P(A \cap B | C) = P(A | C)P(B | C). \]  

Below is an alternative characterization of conditional independence.

**Proposition 5.4.** \( A \) and \( B \) are independent conditional on \( C \) if and only if

\[ P(A | C) = P(A | B \cap C). \]  

**Proof.** Note that

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

Hence by comparing with (134), we obtain the assertion. \( \square \)

In other words, when we know that \( C \) have occurred, \( A \) and \( B \) are conditionally independent when the extra knowledge that \( B \) also have occurred does not change the probability of \( A \).

**Example 5.5.** Flip two independent fair coins. Define three events \( A_1, A_2, \) and \( B \) by

\[ A_1 = \{1st \text{ toss is a head}\} \]  

\[ A_2 = \{2nd \text{ toss is a head}\} \]  

\[ B = \{\text{the two tosses give different results}\} \]

Clearly \( A_1 \) and \( A_2 \) are independent. However, note that

\[ P(A_1 \cap A_2 | B) = P(\emptyset) = 0, \]  

\[ P(A_1 | B) = \frac{P(((H, T)))}{P(((H, T), (T, H)))} = \frac{1}{2}, \]  

\[ P(A_2 | B) = \frac{P(((T, H)))}{P(((H, T), (T, H)))} = \frac{1}{2}. \]

Hence \( A_1 \) and \( A_2 \) are not conditionally independent given \( B \). \( \blacklozenge \)

**Exercise 5.6.** Suppose Bob has a coin of unknown probability of heads, \( \Theta \), from the sample space \( \Omega = \{0.2, 0.9\} \). Alice believes that the two probabilities are equally likely. Bob tosses his coin twice independently. Let \( H_i \) be the event that the \( i \)th toss comes up heads, for \( i = 1, 2 \). Let \( B = \{\Theta = 0.2\} \).

(i) Show that the events \( H_1 \) and \( H_2 \) are independent conditional on \( B \), regardless of Alice’s belief.

(ii) Show that the events \( H_1 \) and \( H_2 \) are not independent under Alice’s belief. Is there any prior for Alice such that \( H_i \)’s are independent?