Chapter 6
An Introduction to Discrete Probability

6.1 Sample Space, Outcomes, Events, Probability

Roughly speaking, probability theory deals with experiments whose outcome are not predictable with certainty. We often call such experiments random experiments. They are subject to chance. Using a mathematical theory of probability, we may be able to calculate the likelihood of some event.

In the introduction to his classical book [1] (first published in 1888), Joseph Bertrand (1822–1900) writes (translated from French to English):

“How dare we talk about the laws of chance (in French: le hasard)? Isn’t chance the antithesis of any law? In rejecting this definition, I will not propose any alternative. On a vaguely defined subject, one can reason with authority. ...”

Of course, Bertrand’s words are supposed to provoke the reader. But it does seem paradoxical that anyone could claim to have a precise theory about chance! It is not my intention to engage in a philosophical discussion about the nature of chance. Instead, I will try to explain how it is possible to build some mathematical tools that can be used to reason rigorously about phenomena that are subject to chance. These tools belong to probability theory. These days, many fields in computer science such as machine learning, cryptography, computational linguistics, computer vision, robotics, and of course algorithms, rely a lot on probability theory. These fields are also a great source of new problems that stimulate the discovery of new methods and new theories in probability theory.

Although this is an oversimplification that ignores many important contributors, one might say that the development of probability theory has gone through four eras whose key figures are: Pierre de Fermat and Blaise Pascal, Pierre–Simon Laplace, and Andrey Kolmogorov. Of course, Gauss should be added to the list; he made major contributions to nearly every area of mathematics and physics during his lifetime. To be fair, Jacob Bernoulli, Abraham de Moivre, Pafnuty Chebyshev, Alexandr Lyapunov, Andrei Markov, Émile Borel, and Paul Lévy should also be added to the list.
Before Kolmogorov, probability theory was a subject that still lacked precise definitions. In 1933, Kolmogorov provided a precise axiomatic approach to probability theory which made it into a rigorous branch of mathematics; with even more applications than before!

The first basic assumption of probability theory is that even if the outcome of an experiment is not known in advance, the set of all possible outcomes of an experiment is known. This set is called the sample space or probability space. Let us begin with a few examples.

**Example 6.1.** If the experiment consists of flipping a coin twice, then the sample space consists of all four strings

\[
\Omega = \{HH, HT, TH, TT\},
\]

where H stands for heads and T stands for tails.

If the experiment consists in flipping a coin five times, then the sample space \(\Omega\) is the set of all strings of length five over the alphabet \(\{H, T\}\), a set of \(2^5 = 32\) strings,

\[
\Omega = \{HHHHH,THHHH,HTHHH,TTHHH,\ldots,TTTTT\}.
\]

**Example 6.2.** If the experiment consists in rolling a pair of dice, then the sample space \(\Omega\) consists of the 36 pairs in the set

\[
\Omega = D \times D
\]

with

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

where the integer \(i\) corresponds to the number on the face of the dice facing up. Here we assume that one dice is rolled first and then another dice is rolled second.

**Example 6.3.** In the game of bridge, the deck has 52 cards and each player receives a hand of 13 cards. Let \(\Omega\) be the sample space of all possible hands. This time it is not possible to enumerate the sample space explicitly. Indeed, there are
Each member of a sample space is called an outcome or an elementary event. Typically, we are interested in experiments consisting of a set of outcomes. For example, in Example 6.1 where we flip a coin five times, the event that exactly one of the coins shows heads is

\[ A = \{HHTTT, THTTT, TTHTT, TTTHT, TTTTH\} \]

The event \( A \) consists of five outcomes. In Example 6.3, the event that we get “doubles” when we roll two dice, namely that each dice shows the same value is,

\[ B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\} \]

an event consisting of 6 outcomes.

The second basic assumption of probability theory is that every elementary event \( \omega \) of a sample space \( \Omega \) is assigned some probability \( \Pr(\omega) \). Intuitively, \( \Pr(\omega) \) is the probability that the elementary event \( \omega \) may occur. It is convenient to normalize probabilities, so we require that

\[ 0 \leq \Pr(\omega) \leq 1. \]

If \( \Omega \) is finite, we also require that

\[ \sum_{\omega \in \Omega} \Pr(\omega) = 1. \]

The function \( \Pr \) is often called a probability distribution on \( \Omega \). Indeed, it distributes the probability of 1 among the events \( \omega \).

In many cases, we assume that the probably distribution is uniform, which means that every elementary event has the same probability.

For example, is we assume that our coins are “fair,” then when we flip a coin five times, since each event in \( \Omega \) is equally likely, the probability of each outcome \( \omega \in \Omega \) is

\[ \Pr(\omega) = \frac{1}{32}. \]
If we assume that our dice are “fair,” namely that each of the six possibilities for a particular dice has probability \( \frac{1}{6} \), then each of the 36 rolls \( \omega \in \Omega \) has probability

\[
\Pr(\omega) = \frac{1}{36}.
\]

We can also consider “loaded dice” in which there is a different distribution of probabilities. For example, let

\[
\begin{align*}
\Pr_1(1) &= \Pr_1(6) = \frac{1}{4}, \\
\Pr_1(2) &= \Pr_1(3) = \Pr_1(4) = \Pr_1(5) = \frac{1}{8}.
\end{align*}
\]

These probabilities add up to 1, so \( \Pr_1 \) is a probability distribution on \( D \). We can assign probabilities to the elements of \( \Omega = D \times D \) by the rule

\[
\Pr_{11}(d, d') = \Pr_1(d)\Pr_1(d').
\]

We can easily check that

\[
\sum_{\omega \in \Omega} \Pr_{11}(\omega) = 1,
\]

so \( \Pr_{11} \) is indeed a probability distribution on \( \Omega \). For example, we get

\[
\Pr_{11}(6, 3) = \Pr_1(6)\Pr_1(3) = \frac{1}{4} \cdot \frac{1}{8} = \frac{1}{32}.
\]

Let us summarize all this with the following definition.

**Definition 6.1.** A **finite discrete probability space** (or **finite discrete sample space**) is a finite set \( W \) of outcomes or elementary events \( \omega \in W \), together with a function \( \Pr: \Omega \to \mathbb{R} \), called **probability measure** (or **probability distribution**) satisfying the following properties:

\[
0 \leq \Pr(\omega) \leq 1 \quad \text{for all } \omega \in \Omega.
\]

\[
\sum_{\omega \in \Omega} \Pr(\omega) = 1.
\]

An **event** is any subset \( A \) of \( \Omega \). The probability of an even \( A \) is defined as

\[
\Pr(A) = \sum_{\omega \in A} \Pr(\omega).
\]

Definition 6.1 immediately implies that

\[
\Pr(\emptyset) = 0 \quad \Pr(\Omega) = 1.
\]
For another example, if we consider the event

\[ A = \{\text{HTTTT, THTTT, TTHTT, TTTHT, TTTTH}\} \]

that in flipping a coin five times, heads turns up exactly once, the probability of this event is

\[ \Pr(A) = \frac{5}{32}. \]

If we use the probability distribution \( \Pr \) on the sample space \( \Omega \) of pairs of dice, the probability of the event of having doubles

\[ B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}, \]

is

\[ \Pr(B) = 6 \cdot \frac{1}{36} = \frac{1}{6}. \]

However, using the probability distribution \( \Pr_{11} \), we obtain

\[ \Pr(B) = \frac{1}{16} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{16} = \frac{3}{16} > \frac{1}{16}. \]

Loading the dice makes the event “having doubles” more probable.

It should be noted that a definition slightly more general than Definition 6.1 is needed if we want to allow \( \Omega \) to be infinite. In this case, the following definition is used.

**Definition 6.2.** A discrete probability space (or discrete sample space) is a triple \((\Omega, \mathcal{F}, \Pr)\) consisting of:

1. A nonempty countably infinite set \( \Omega \) of outcomes or elementary events.
2. The set \( \mathcal{F} \) of all subsets of \( \Omega \), called the set of events.
3. A function \( \Pr: \mathcal{F} \to \mathbb{R} \), called probability measure (or probability distribution) satisfying the following properties:
   a. \( 0 \leq \Pr(A) \leq 1 \) for all \( A \in \mathcal{F} \).
   b. \( \Pr(\Omega) = 1 \).
   c. For any sequence of pairwise disjoint events \( E_1, E_2, \ldots, E_i, \ldots \) in \( \mathcal{F} \) (which means that \( E_i \cap E_j = \emptyset \) for all \( i \neq j \)), we have

\[ \Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i). \]

The main thing to observe is that \( \Pr \) is now defined directly on events, since events may be infinite. The third axiom of a probability measure implies that
\[ \Pr(\emptyset) = 0. \]

It may also be desirable to assume that the family \( \mathcal{F} \) of events is a proper subset of the power set of \( \Omega \). In this case, \( \mathcal{F} \) is called the family of measurable events, and \( \mathcal{F} \) has certain closure properties that make it a \( \sigma \)-algebra (also called a \( \sigma \)-field). We may even allow \( \Omega \) to be countably infinite. In this case, we drop the word discrete from discrete probability space.

**Remark:** A \( \sigma \)-algebra is a nonempty family \( \mathcal{F} \) of subsets of \( \Omega \) satisfying the following properties:

1. \( \emptyset \in \mathcal{F} \).
2. For every subset \( A \subseteq \Omega \), if \( A \in \mathcal{F} \) then \( \overline{A} \in \mathcal{F} \).
3. For every countable family \( (A_i)_{i \geq 1} \) of subsets \( A_i \in \mathcal{F} \), we have \( \bigcup_{i \geq 1} A_i \in \mathcal{F} \).

Note that every \( \sigma \)-algebra is a Boolean algebra (see Section 7.11, Definition 7.14), but the closure property (3) is very strong and adds spice to the story.

In this chapter, we will deal mostly with finite discrete probability spaces, and occasionally with discrete probability spaces with a countably infinite sample space. In this latter case, we always assume that \( \mathcal{F} = 2^\Omega \), and we omit it (that is, we write \( (\Omega, \Pr) \) instead of \( (\Omega, \mathcal{F}, \Pr) \)).

Because events are subsets of the sample space \( \Omega \), they can be combined using the set operations, union, intersection, and complementation. If the sample space \( \Omega \) is finite, the definition for the probability \( \Pr(A) \) of an event \( A \subseteq \Omega \) given in Definition 6.1 shows that if \( A, B \) are two disjoint events (this means that \( A \cap B = \emptyset \)), then

\[ \Pr(A \cup B) = \Pr(A) + \Pr(B). \]

More generally, if \( A_1, \ldots, A_n \) are any pairwise disjoint events, then

\[ \Pr(A_1 \cup \cdots \cup A_n) = \Pr(A_1) + \cdots + \Pr(A_n). \]

It is natural to ask whether the probabilities \( \Pr(A \cup B) \), \( \Pr(A \cap B) \) and \( \Pr(\overline{A}) \) can be expressed in terms of \( \Pr(A) \) and \( \Pr(B) \), for any two events \( A, B \in \Omega \). In the first and the third case, we have the following simple answer.

**Proposition 6.1.** Given any (finite) discrete probability space \( (\Omega, \Pr) \), for any two events \( A, B \subseteq \Omega \), we have

\[ \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \]

\[ \Pr(\overline{A}) = 1 - \Pr(A). \]

Furthermore, if \( A \subseteq B \), then \( \Pr(A) \subseteq \Pr(B) \).

**Proof.** Observe that we can write \( A \cup B \) as the following union of pairwise disjoint subsets:

\[ A \cup B = (A \cap B) \cup (A - B) \cup (B - A). \]
Then, using the observation made just before Proposition 6.1, since we have the disjoint unions \( A = (A \cap B) \cup (A - B) \) and \( B = (A \cap B) \cup (B - A) \), using the disjointness of the various subsets, we have

\[
\begin{align*}
Pr(A \cup B) &= Pr(A \cap B) + Pr(A - B) + Pr(B - A) \\
Pr(A) &= Pr(A \cap B) + Pr(A - B) \\
Pr(B) &= Pr(A \cap B) + Pr(B - A),
\end{align*}
\]

and from these we obtain

\[
Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B).
\]

The equation \( Pr(\overline{A}) = 1 - Pr(A) \) follows from the fact that \( A \cap \overline{A} = \emptyset \) and \( A \cup \overline{A} = \Omega \), so

\[
1 = Pr(\Omega) = Pr(A) + Pr(\overline{A}).
\]

If \( A \subseteq B \), then \( A \cap B = A \), so \( B = (A \cap B) \cup (B - A) = A \cup (B - A) \), and since \( A \) and \( B - A \) are disjoint, we get

\[
Pr(B) = Pr(A) + Pr(B - A).
\]

Since probabilities are nonnegative, the above implies that \( Pr(A) \leq Pr(B) \). \( \square \)

**Remark:** Proposition 6.1 still holds when \( \Omega \) is infinite as a consequence of axioms (a)–(c) of a probability measure. Also, the equation

\[
Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)
\]

can be generalized to any sequence of \( n \) events. In fact, we already showed this as the Principle of Inclusion–Exclusion, Version 2 (Theorem 5.2).

In general, the probability \( Pr(A \cap B) \) of the event \( A \cap B \) cannot be expressed in a simple way in terms of \( Pr(A) \) and \( Pr(B) \). However, in many cases we observe that \( Pr(A \cap B) = Pr(A)Pr(B) \). If this holds, we say that \( A \) and \( B \) are independent.

**Definition 6.3.** Given a discrete probability space \((\Omega, Pr)\), two events \( A \) and \( B \) are independent if

\[
Pr(A \cap B) = Pr(A)Pr(B).
\]

Two events are dependent if they are not independent.

For example, in the sample space of 5 coin flips, we have the events

\[
A = \{HHw \mid w \in \{H,T\}^3\} \cup \{HTw \mid w \in \{H,T\}^3\},
\]

the event in which the first flip is \( H \), and

\[
B = \{HHw \mid w \in \{H,T\}^3\} \cup \{THw \mid w \in \{H,T\}^3\},
\]
the event in which the second flip is H. Since $A$ and $B$ contain 16 outcomes, we have

$$\Pr(A) = \Pr(B) = \frac{16}{32} = \frac{1}{2}.$$ 

The intersection of $A$ and $B$ is

$$A \cap B = \{ HHw \mid w \in \{H,T\}^3 \},$$

the event in which the first two flips are H, and since $A \cap B$ contains 8 outcomes, we have

$$\Pr(A \cap B) = \frac{8}{32} = \frac{1}{4}.$$ 

Since

$$\Pr(A \cap B) = \frac{1}{4},$$

and

$$\Pr(A)\Pr(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

we see that $A$ and $B$ are independent events. On the other hand, if we consider the events

$$A = \{ TTTTT, HHTTT \}$$

and

$$B = \{ TTTTT, HTTTT \},$$

we have

$$\Pr(A) = \Pr(B) = \frac{2}{32} = \frac{1}{16},$$

and since

$$A \cap B = \{ TTTTT \},$$

we have

$$\Pr(A \cap B) = \frac{1}{32}.$$ 

It follows that

$$\Pr(A)\Pr(B) = \frac{1}{16} \cdot \frac{1}{16} = \frac{1}{256},$$

but

$$\Pr(A \cap B) = \frac{1}{32},$$

so $A$ and $B$ are not independent.

**Example 6.4.** We close this section with a classical problem in probability known as the *birthday problem*. Consider $n < 365$ individuals and assume for simplicity that nobody was born on February 29. In this problem, the sample space is the set of all $365^n$ possible choices of birthdays for $n$ individuals, and let us assume that they are all equally likely. This is equivalent to assuming that each of the 356 days of the year is an equally likely birthday for each individual, and that the assignments
of birthdays to distinct people are independent. Note that this does not take twins into account! What is the probability that two (or more) individuals have the same birthday?

To solve this problem, it is easier to compute the probability that no two individuals have the same birthday. We can choose \( \binom{365}{n} \) distinct birthdays in \( 365^n \) ways, and these can be assigned to \( n \) people in \( n! \) ways, so there are

\[
\binom{365}{n} n! = 365 \cdot 364 \cdots (365 - n + 1)
\]

configurations where no two people have the same birthday. There are \( 365^n \) possible choices of birthdays, so the probability that no two people have the same birthday is

\[
q = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n} = \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n - 1}{365}\right),
\]

and thus, the probability that two people have the same birthday is

\[
p = 1 - q = 1 - \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n - 1}{365}\right).
\]

In the proof of Proposition 5.15, we showed that \( x \leq e^{x-1} \) for all \( x \in \mathbb{R} \), so \( 1 - x \leq e^{-x} \) for all \( x \in \mathbb{R} \), and we can bound \( q \) as follows:

\[
q = \prod_{i=1}^{n-1} \left(1 - \frac{i}{365}\right)
\]

\[
q \leq \prod_{i=1}^{n-1} e^{-i/365}
= e^{-\sum_{i=1}^{n-1} \frac{i}{365}}
= e^{-\frac{n(n-1)}{2 \cdot 365}}.
\]

If we want the probability \( q \) that no two people have the same birthday to be at most 1/2, it suffices to require

\[
e^{-\frac{n(n-1)}{2 \cdot 365}} \leq \frac{1}{2},
\]

that is, \( -n(n - 1)/(2 \cdot 365) \leq \ln(1/2) \), which can be written as

\[
n(n - 1) \geq 2 \cdot 365 \ln 2.
\]

The roots of the quadratic equation

\[
n^2 - n - 2 \cdot 365 \ln 2 = 0
\]

are
\[ m = \frac{1 \pm \sqrt{1 + 8 \cdot 365 \ln 2}}{2}, \]

and we find that the positive root is approximately \( m = 23 \). In fact, we find that if \( n = 23 \), then \( p \approx 50.7\% \). If \( n = 30 \), we calculate that \( p \approx 71\% \).

What if we want at least three people to share the same birthday? Then \( n = 88 \) does it, but this is harder to prove! See Ross [12], Section 3.4.

Next, we define what is perhaps the most important concept in probability: that of a random variable.

### 6.2 Random Variables and their Distributions

In many situations, given some probability space \((\Omega, Pr)\), we are more interested in the behavior of functions \( X : \Omega \to \mathbb{R} \) defined on the sample space \( \Omega \) than in the probability space itself. Such functions are traditionally called random variables, a somewhat unfortunate terminology since these are functions. Now, given any real number \( a \), the inverse image of \( a \)

\[ X^{-1}(a) = \{ \omega \in \Omega \mid X(\omega) = a \}, \]

is a subset of \( \Omega \), thus an event, so we may consider the probability \( Pr(X^{-1}(a)) \), denoted (somewhat improperly) by

\[ Pr(X = a). \]

This function of \( a \) is of great interest, and in many cases it is the function that we wish to study. Let us give a few examples.

**Example 6.5.** Consider the sample space of 5 coin flips, with the uniform probability measure (every outcome has the same probability \( 1/32 \)). Then, the number of times \( X(\omega) \) that H appears in the sequence \( \omega \) is a random variable. We determine that

\[
\begin{align*}
Pr(X = 0) &= \frac{1}{32} \\
Pr(X = 1) &= \frac{5}{32} \\
Pr(X = 2) &= \frac{10}{32} \\
Pr(X = 3) &= \frac{10}{32} \\
Pr(X = 4) &= \frac{5}{32} \\
Pr(X = 5) &= \frac{1}{32}
\end{align*}
\]

The function defined \( Y \) such that \( Y(\omega) = 1 \) if \( H \) appears in \( \omega \), and \( Y(\omega) = 0 \) otherwise, is a random variable. We have

\[
\begin{align*}
Pr(Y = 0) &= \frac{1}{2} \\
Pr(Y = 1) &= \frac{1}{2}
\end{align*}
\]
Example 6.6. Let \( \Omega = D \times D \) be the sample space of dice rolls, with the uniform probability measure \( \Pr \) (every outcome has the same probability \( 1/36 \)). The sum \( S(\omega) \) of the numbers on the two dice is a random variable. For example,

\[
S(2,5) = 7.
\]

The value of \( S \) is any integer between 2 and 12, and if we compute \( \Pr(S = s) \) for \( s = 2, \ldots, 12 \), we find the following table:

<table>
<thead>
<tr>
<th>( s )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr(S = s) )</td>
<td>( \frac{1}{36} )</td>
<td>( \frac{2}{36} )</td>
<td>( \frac{3}{36} )</td>
<td>( \frac{4}{36} )</td>
<td>( \frac{5}{36} )</td>
<td>( \frac{6}{36} )</td>
<td>( \frac{5}{36} )</td>
<td>( \frac{4}{36} )</td>
<td>( \frac{3}{36} )</td>
<td>( \frac{2}{36} )</td>
<td>( \frac{1}{36} )</td>
</tr>
</tbody>
</table>

Here is a “real” example from computer science.

Example 6.7. Our goal is to sort a sequence \( S = (x_1, \ldots, x_n) \) of \( n \) distinct real number in increasing order. We use a recursive method known as quicksort which proceeds as follows:

1. If \( S \) has one or zero elements return \( S \).
2. Pick some element \( x = x_i \) in \( S \) called the pivot.
3. Reorder \( S \) in such a way that for every number \( x_j \neq x \) in \( S \), if \( x_j < x \), then \( x_j \) is moved to a list \( S_1 \), else if \( x_j > x \) then \( x_j \) is moved to a list \( S_2 \).
4. Apply this algorithm recursively to the list of elements in \( S_1 \) and to the list of elements in \( S_2 \).
5. Return the sorted list \( S_1 \cup x \cup S_2 \).

Let us run the algorithm on the input list

\[
S = (1, 5, 9, 2, 3, 8, 7, 14, 12, 10).
\]

We can represent the choice of pivots and the steps of the algorithm by an ordered binary tree as shown in Figure 6.3. Except for the root node, every node corresponds to the choice of a pivot, say \( x \). The list \( S_1 \) is shown as a label on the left of node \( x \), and the list \( S_2 \) is shown as a label on the right of node \( x \). A leaf node is a node such that \( |S_1| \leq 1 \) and \( |S_2| \leq 1 \). If \( |S_1| \geq 2 \), then \( x \) has a left child, and if \( |S_2| \geq 2 \), then \( x \) has a right child. Let us call such a tree a computation tree. Observe that except for minor cosmetic differences, it is a binary search tree. The sorted list can be retrieved by a suitable traversal of the computation tree, and is

\[
(1, 2, 3, 5, 7, 8, 9, 10, 12, 14).
\]

If you run this algorithm on a few more examples, you will realize that the choice of pivots greatly influences how many comparisons are needed. If the pivot is chosen at each step so that the size of the lists \( S_1 \) and \( S_2 \) is roughly the same, then the number of comparisons is small compared to \( n \), in fact \( O(n \ln n) \). On the other hand, with a poor choice of pivot, the number of comparisons can be as bad as \( n(n-1)/2 \).

In order to have a good “average performance,” one can randomize this algorithm by assuming that each pivot is chosen at random. What this means is that whenever
it is necessary to pick a pivot from some list $Y$, some procedure is called and this procedure returns some element chosen at random from $Y$. How exactly this done is an interesting topic in itself but we will not go into this. Let us just say that the pivot can be produced by a random number generator, or by spinning a wheel containing the numbers in $Y$ on it, or by a rolling a dice with as many faces as the numbers in $Y$! What we do assume is that the probability distribution that a number is chosen from a list $Y$ is uniform, and that successive choices of pivots are independent. How do we model this as a probability space?

Here is a way to do it. Use the computation trees defined above! Simply add to every edge the probability that one of the elements of the corresponding list, say $Y$, was chosen uniformly, namely $1/|Y|$. So, given an input list $S$ of length $n$, the sample space $\Omega$ is the set of all computation trees $T$ with root label $S$. We assign a probability to the trees $T$ in $\Omega$ as follows: If $n = 0$, then there is a single tree and its probability is 1. If $n \geq 2$, for every leaf of $T$, multiply the probabilities along the path from the root to that leaf and then add up the probabilities assigned to these leaves. This is $\Pr(T)$. We leave it as an exercise to prove that the sum of the probabilities of all the trees in $\Omega$ is equal to 1.

A random variable of great interest on $(\Omega, \Pr)$ is the number $X$ of comparisons performed by the algorithm. To analyse the average running time of this algorithm, it is necessary to determine when the first (or the last) element of a sequence

$$Y = (y_i, \ldots, y_j)$$

is chosen as a pivot. It can be shown that this probability is equal to
To carry out the analysis further requires the notion of expectation that has not yet been defined. See Example 6.23 for a complete analysis.

Let us now give an official definition of a random variable.

**Definition 6.4.** Given a (finite) discrete probability space \((\Omega, \Pr)\), a random variable is any function \(X: \Omega \to \mathbb{R}\). For any real number \(a \in \mathbb{R}\), we define \(\Pr(X = a)\) as the probability

\[
\Pr(X = a) = \Pr(X^{-1}(\{a\})) = \Pr(\{\omega \in \Omega \mid X(\omega) = a\}),
\]

and \(\Pr(X \leq a)\) as the probability

\[
\Pr(X \leq a) = \Pr(X^{-1}((-\infty, a])) = \Pr(\{\omega \in \Omega \mid X(\omega) \leq a\}).
\]

The function \(f: \mathbb{R} \to [0, 1]\) given by

\[
f(a) = \Pr(X = a), \quad a \in \mathbb{R}
\]

is the probability mass function of \(X\), and the function \(F: \mathbb{R} \to [0, 1]\) given by

\[
F(a) = \Pr(X \leq a), \quad a \in \mathbb{R}
\]

is the cumulative distribution function of \(X\).

The term probability mass function is abbreviated as *p.m.f*, and cumulative distribution function is abbreviated as *c.d.f*. It is unfortunate and confusing that both the probability mass function and the cumulative distribution function are often abbreviated as *distribution function*.

The probability mass function \(f\) for the sum \(S\) of the numbers on two dice from Example 6.6 is shown in Figure 6.4, and the corresponding cumulative distribution function \(F\) is shown in Figure 6.5.

![Fig. 6.4 The probability mass function for the sum of the numbers on two dice](chart.png)
If $\Omega$ is finite, then $f$ only takes finitely many nonzero values; it is it very discontinuous! The c.d.f $F$ of $S$ shown in Figure 6.5 has jumps (steps). Observe that the size of the jump at every value $a$ is equal to $f(a) = \Pr(S = a)$.

The cumulative distribution function $F$ is monotonic nondecreasing, which means that if $a \leq b$, then $F(a) \leq F(b)$. It is piecewise constant with jumps, but it is right-continuous, which means that $\lim_{h>0,h\to0}F(a+h) = F(a)$. For any $a \in \mathbb{R}$, because $F$ is nondecreasing, we can define $F(a-)$ by

$$F(a-) = \lim_{h\downarrow0}F(a-h) = \lim_{h>0,h\to0}F(a-h).$$

These properties are clearly illustrated by the c.d.f on Figure 6.5.

The functions $f$ and $F$ determine each other, because given the probability mass function $f$, the function $F$ is defined by

$$F(a) = \sum_{x \leq a} f(x),$$

and given the cumulative distribution function $F$, the function $f$ is defined by

$$f(a) = F(a) - F(a-).$$

If the sample space $\Omega$ is countably infinite, then $f$ and $F$ are still defined as above but in
\[ F(a) = \sum_{x \leq a} f(x), \]

the expression on the righthand side is the limit of an infinite sum (of positive terms).

**Remark:** If \( \Omega \) is not countably infinite, then we are dealing with a probability space \((\Omega, \mathcal{F}, \Pr)\) where \( \mathcal{F} \) may be a proper subset of \(2^\Omega\), and in Definition 6.4, we need the extra condition that a random variable is a function \( X : \Omega \to \mathbb{R} \) such that \( X^{-1}(a) \in \mathcal{F} \) for all \( a \in \mathbb{R} \). (The function \( X \) needs to be \( \mathcal{F} \)-measurable.) In this more general situation, it is still true that

\[ f(a) = \Pr(X = a) = F(a) - F(a^-), \]

but \( F \) cannot generally be recovered from \( f \). If the c.d.f \( F \) of a random variable \( X \) can be expressed as

\[ F(x) = \int_{-\infty}^{x} f(y)dy, \]

for some nonnegative (Lebesgue) integrable function \( f \), then we say that \( F \) and \( X \) are absolutely continuous (please, don’t ask me what type of integral!). The function \( f \) is called a probability density function of \( X \) (for short, p.d.f).

In this case, \( F \) is continuous, but more is true. The function \( F \) is uniformly continuous, and it is differentiable almost everywhere, which means that the set of input values for which it is not differentiable is a set of (Lebesgue) measure zero. Furthermore, \( F' = f \) almost everywhere.

Random variables whose distributions can be expressed as above in terms of a density function are often called continuous random variables. In contrast with the discrete case, if \( X \) is a continuous random variable, then

\[ \Pr(X = x) = 0 \quad \text{for all } x \in \mathbb{R}. \]

As a consequence, some of the definitions given in the discrete case in terms of the probabilities \( \Pr(X = x) \) need to be modified; replacing \( \Pr(X = x) \) by \( \Pr(X \leq x) \) usually works.

In the general case where the cdf \( F \) of a random variable \( X \) has discontinuities, we say that \( X \) is a discrete random variable if \( X(\omega) \neq 0 \) for at most countably many \( \omega \in \Omega \). Equivalently, the image of \( X \) is finite or countably infinite. In this case, the mass function of \( X \) is well defined, and it can be viewed as a discrete version of a density function.

In the discrete setting where the sample space \( \Omega \) is finite, it is usually more convenient to use the probability mass function \( f \), and to abuse language and call it the distribution of \( X \).

**Example 6.8.** Suppose we flip a coin \( n \) times, but this time, the coin is not necessarily fair, so the probability of landing heads is \( p \) and the probability of landing tails is \( 1 - p \). The sample space \( \Omega \) is the set of strings of length \( n \) over the alphabet \( \{H,T\} \). Assume that the coin flips are independent, so that the probability of an event \( \omega \in \Omega \) is obtained by replacing \( H \) by \( p \) and \( T \) by \( 1 - p \) in \( \omega \). Then, let \( X \) be
the random variable defined such that \( X(\omega) \) is the number of heads in \( \omega \). For any \( i \) with \( 0 \leq i \leq n \), since there are \( \binom{n}{i} \) subsets with \( i \) elements, and since the probability of a sequence \( \omega \) with \( i \) occurrences of H is \( p^i(1-p)^{n-i} \), we see that the distribution of \( X \) (mass function) is given by

\[
f(i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, \ldots, n,
\]

and 0 otherwise. This is an example of a \textit{binomial distribution}.

\textbf{Example 6.9.} As in Example 6.8, assume that we flip a biased coin, where the probability of landing heads is \( p \) and the probability of landing tails is \( 1 - p \). However, this time, we flip our coin any finite number of times (not a fixed number), and we are interested in the event that heads first turns up. The sample space \( \Omega \) is the infinite set of strings over the alphabet \{H, T\} of the form

\[
\Omega = \{H, TH, TTH, \ldots, T^nH, \ldots\}.
\]

Assume that the coin flips are independent, so that the probability of an event \( \omega \in \Omega \) is obtained by replacing H by \( p \) and T by \( 1 - p \) in \( \omega \). Then, let \( X \) be the random variable defined such that \( X(\omega) = n \) iff \( |\omega| = n \), else 0. In other words, \( X \) is the number of trials until we obtain a success. Then, it is clear that

\[
f(n) = (1-p)^{n-1} p, \quad n \geq 1.
\]

and 0 otherwise. This is an example of a \textit{geometric distribution}.

The process in which we flip a coin \( n \) times is an example of a process in which we perform \( n \) independent trials, each of which results in success or failure (such trials that result exactly two outcomes, success or failure, are known as \textit{Bernoulli trials}). Such processes are named after Jacob Bernoulli, a very significant contributor to probability theory after Fermat and Pascal.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig66.png}
\caption{Jacob (Jacques) Bernoulli (1654–1705)}
\end{figure}

\textbf{Example 6.10.} Let us go back to Example 6.8, but assume that \( n \) is large and that the probability \( p \) of success is small, which means that we can write \( np = \lambda \) with \( \lambda \)
of “moderate” size. Let us show that we can approximate the distribution \( f \) of \( X \) in an interesting way. Indeed, we can write

\[
f(i) = \binom{n}{i} p^i (1-p)^{n-i}
\]

\[
= \frac{n!}{i!(n-i)!} \left( \frac{\lambda}{n} \right)^i \left( 1 - \frac{\lambda}{n} \right)^{n-i}
\]

\[
= \frac{n(n-1) \cdots (n-i+1) \lambda^i}{n^i} \frac{1}{i!} \left( 1 - \frac{\lambda}{n} \right)^{n} \left( 1 - \frac{\lambda}{n} \right)^{-i}.
\]

Now, for \( n \) large and \( \lambda \) moderate, we have

\[
\left( 1 - \frac{\lambda}{n} \right)^n \approx e^{-\lambda} \quad \left( 1 - \frac{\lambda}{n} \right)^{-i} \approx 1 \quad \frac{n(n-1) \cdots (n-i+1)}{n^i} \approx 1,
\]

so we obtain

\[
f(i) \approx e^{-\lambda} \frac{\lambda^i}{i!}.
\]

The above is called a Poisson distribution with parameter \( \lambda \). It is named after the French mathematician Denis Poisson.

It turns out that quite a few random variables occurring in real life obey the Poisson probability law (by this, we mean that their distribution is the Poisson distribution). Here are a few examples:

1. The number of misprints on a page (or a group of pages) in a book.
2. The number of people in a community whose age is over a hundred.
3. The number of wrong telephone numbers that are dialed in a day.
4. The number of customers entering a post office each day.
5. The number of vacancies occurring in a year in the federal judicial system.

As we will see later on, the Poisson distribution has some nice mathematical properties, and the so-called Poisson paradigm which consists in approximating the distribution of some process by a Poisson distribution is quite useful.
6.3 Conditional Probability and Independence

In general, the occurrence of some event $B$ changes the probability that another event $A$ occurs. It is then natural to consider the probability denoted $\Pr(A \mid B)$ that if an event $B$ occurs, then $A$ occurs. As in logic, if $B$ does not occur not much can be said, so we assume that $\Pr(B) \neq 0$.

**Definition 6.5.** Given a discrete probability space $(\Omega, \Pr)$, for any two events $A$ and $B$, if $\Pr(B) \neq 0$, then we define the **conditional probability** $\Pr(A \mid B)$ that $A$ occurs given that $B$ occurs as

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$  

**Example 6.11.** Suppose we roll two fair dice. What is the conditional probability that the sum of the numbers on the dice exceeds 6, given that the first shows 3? To solve this problem, let

$$B = \{(3, j) \mid 1 \leq j \leq 6\}$$

be the event that the first dice shows 3, and

$$A = \{(i, j) \mid i+j \geq 7, 1 \leq i, j \leq 6\}$$

be the event that the total exceeds 6. We have

$$A \cap B = \{(3, 4), (3, 5), (3, 6)\},$$

so we get

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{3}{36} = \frac{1}{2}.$$ 

The next example is perhaps a little more surprising.

**Example 6.12.** A family has two children. What is the probability that both are boys, given at least one is a boy?

There are four possible combinations of sexes, so the sample space is

$$\Omega = \{GG, GB, BG, BB\},$$

and we assume a uniform probability measure (each outcome has probability $1/4$).

Introduce the events

$$B = \{GB, BG, BB\}$$

of having at least one boy, and

$$A = \{BB\}$$

of having two boys. We get

$$A \cap B = \{BB\},$$

and so
6.3 Conditional Probability and Independence

\[ Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{1}{4} \cdot \frac{3}{4} = \frac{1}{3}. \]

Contrary to the popular belief that \( Pr(A \mid B) = \frac{1}{2} \), it is actually equal to \( \frac{1}{3} \). Now, if the question is what is the probability that both are boys, given that the first child is a boy, then the answer is indeed \( \frac{1}{2} \).

The next example is known as the “Monty Hall Problem,” a standard example of every introduction to probability theory.

Example 6.13. On the old television game *Let’s make a deal*, a contestant is presented with a choice of three (closed) doors. Behind exactly one door is a terrific prize. The other doors conceal cheap items. First, the contestant is asked to choose a door. Then, the host of the show (Monty Hall) shows the contestant one of the worthless prizes behind one of the other doors. At this point, there are two closed doors, and the contestant is given the opportunity to switch from his original choice to the other closed door. The question is, is it better for the contestant to stick to his original choice or to switch door?

We can analyze this problem using conditional probabilities. Without loss of generality, assume that the contestant chooses door 1. If the prize is actually behind door 1, then the host will show door 2 or door 3 with equal probability \( \frac{1}{2} \). However, if the prize is behind door 2, then the host will open door 3 with probability 1, and if the prize is behind door 3, then the host will open door 2 with probability 1. Write \( P_i \) for “the prize is behind door \( i \),” with \( i = 1, 2, 3 \), and \( D_j \) for “the host opens door \( D_j \),” for \( j = 2, 3 \). Here, it is not necessary to consider the choice \( D_1 \) since a sensible host will never open door 1. We can represent the sequences of choices occurring in the game by a tree known as probability tree or tree of possibilities, shown in Figure 6.8.

Every leaf corresponds to a path associated with an elementary event, so the sample space is

\[ \Omega = \{P_1;D_2,P_1;D_3,P_2;D_3,P_3;D_2\}. \]

The probability of an elementary event is obtained by multiplying the probabilities along the corresponding path, so we have

\[ Pr(P_1;D_2) = \frac{1}{6}, \quad Pr(P_1;D_3) = \frac{1}{6}, \quad Pr(P_2;D_3) = \frac{1}{3}, \quad Pr(P_3;D_2) = \frac{1}{3}. \]

Suppose that the host reveals door 2. What should the contestant do?

The events of interest are:

1. The prize is behind door 1; that is, \( A = \{P_1;D_2,P_1;D_3\} \).
2. The prize is behind door 3; that is, \( B = \{P_3;D_2\} \).
3. The host reveals door 2; that is, \( C = \{P_1;D_2,P_3;D_2\} \).

Whether or not the contestant should switch door depends on the values of the conditional probabilities

1. \( Pr(A \mid C) \): the prize is behind door 1, given that the host reveals door 2.
Fig. 6.8 The tree of possibilities in the Monty Hall problem

2. \( \Pr(B \mid C) \): the prize is behind door 3, given that the host reveals door 2.

We have \( A \cap C = \{P1; D2\} \), so

\[
\Pr(A \cap C) = 1/6,
\]

and

\[
\Pr(C) = \Pr(\{P1; D2, P3; D2\}) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2},
\]

so

\[
\Pr(A \mid C) = \frac{\Pr(A \cap C)}{\Pr(C)} = \frac{1/6}{1/2} = \frac{1}{3}.
\]

We also have \( B \cap C = \{P3; D2\} \), so

\[
\Pr(B \cap C) = 1/3,
\]

and

\[
\Pr(B \mid C) = \frac{\Pr(B \cap C)}{\Pr(C)} = \frac{1/3}{1/2} = \frac{2}{3}.
\]

Since \( 2/3 > 1/3 \), the contestant has a greater chance (twice as big) to win the bigger prize by switching door. The same probabilities are derived if the host had revealed door 3.

A careful analysis showed that the contestant has a greater chance (twice as large) of winning big if she/he decides to switch door. Most people say “on intuition” that it is preferable to stick to the original choice, because once one door is revealed, the probability that the valuable prize is behind either of two remaining doors is
1/2. This is incorrect because the door the host opens depends on which door the contestant originally chose.

Let us conclude by stressing that probability trees (trees of possibilities) are very useful in analyzing problems in which sequences of choices involving various probabilities are made.

The next proposition shows various useful formulae due to Bayes.

**Proposition 6.2.** (Bayes’ Rules) For any two events \(A, B\) with \(\Pr(A) > 0\) and \(\Pr(B) > 0\), we have the following formulae:

1. (Bayes’ rule of retrodiction)
   \[
   \Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}.
   \]

2. (Bayes’ rule of exclusive and exhaustive clauses) If we also have \(\Pr(A) < 1\) and \(\Pr(B) < 1\), then
   \[
   \Pr(A) = \Pr(A \mid B)\Pr(B) + \Pr(A \mid \overline{B})\Pr(\overline{B}).
   \]

   More generally, if \(B_1, \ldots, B_n\) form a partition of \(\Omega\) with \(\Pr(B_i) > 0\) \((n \geq 2)\), then
   \[
   \Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i).
   \]

3. (Bayes’ sequential formula) For any sequence of events \(A_1, \ldots, A_n\), we have
   \[
   \Pr\left(\bigcap_{i=1}^{n} A_i\right) = \Pr(A_1)\Pr(A_2 \mid A_1)\Pr(A_3 \mid A_1 \cap A_2) \cdots \Pr\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right).
   \]

**Proof.** The first formula is obvious by definition of a conditional probability. For the second formula, observe that we have the disjoint union
   \[ A = (A \cap B) \cup (A \cap \overline{B}), \]

so
   \[
   \Pr(A) = \Pr(A \cap B) \cup \Pr(A \cap \overline{B})
   = \Pr(A \mid B)\Pr(A) + \Pr(A \mid \overline{B})\Pr(\overline{B}).
   \]

We leave the more general rule as an exercise, and the last rule follows by unfolding definitions.

It is often useful to combine (1) and (2) into the rule
   \[
   \Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid \overline{B})\Pr(\overline{B})}.
   \]
also known as Bayes’ law.

**Example 6.14.** Doctors apply a medical test for a certain rare disease that has the property that if the patient is affected by the disease, then the test is positive in 99% of the cases. However, it happens in 2% of the cases that a healthy patient tests positive. Statistical data shows that one person out of 1000 has the disease. What is the probability for a patient with a positive test to be affected by the disease?

Let $S$ be the event that the patient has the disease, and $+$ and $-$ the events that the test is positive or negative. We know that

$$
\Pr(S) = 0.001 \\
\Pr(\ + \mid S) = 0.99 \\
\Pr(\ + \mid \neg S) = 0.02,
$$

and we have to compute $\Pr(S \mid +)$. We use the rule

$$
\Pr(S \mid +) = \frac{\Pr(\ + \mid S)\Pr(S)}{\Pr(\ + )}.
$$

We also have

$$
\Pr(\ + ) = \Pr(\ + \mid S)\Pr(S) + \Pr(\ + \mid \neg S)\Pr(\neg S),
$$

so we obtain

$$
\Pr(S \mid +) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.02 \times 0.999} \approx \frac{1}{20} = 5\%.
$$

This probability is small but not so small that it instills a lot of confidence in the test! The solution is to apply a better test, but only to all positive patients. Only a small portion of the population will be given that second test because

$$
\Pr(\ + ) = 0.99 \times 0.001 + 0.02 \times 0.999 \approx 0.003.
$$

Redo the calculations with the new data

$$
\Pr(S) = 0.0001 \\
\Pr(\ + \mid S) = 0.99 \\
\Pr(\ + \mid \neg S) = 0.01.
$$

You will find that probability $\Pr(S \mid +)$ is quite small, so the chance of being sick is rather small.

Recall that in Definition 6.3, we defined two events as being independent if

$$
\Pr(A \cap B) = \Pr(A)\Pr(B).
$$

Assuming that $\Pr(A) \neq 0$ and $\Pr(B) \neq 0$, we have
\[
\Pr(A \cap B) = \Pr(A \mid B) \Pr(B) = \Pr(B \mid A) \Pr(A),
\]
so we get the following proposition.

**Proposition 6.3.** For any two events \(A, B\) such that \(\Pr(A) \neq 0\) and \(\Pr(B) \neq 0\), the following statements are equivalent:

1. \(\Pr(A \cap B) = \Pr(A) \Pr(B)\); that is, \(A\) and \(B\) are independent.
2. \(\Pr(B \mid A) = \Pr(B)\).
3. \(\Pr(A \mid B) = \Pr(A)\).

The examples where we flip a coin \(n\) times or roll two dice \(n\) times are examples of independent repeated trials. They suggest the following definition.

**Definition 6.6.** Given two discrete probability spaces \((\Omega_1, \mathcal{F}_1, \Pr_1)\) and \((\Omega_2, \mathcal{F}_2, \Pr_2)\), we define their **product space** as the probability space \((\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \Pr)\), where \(\Pr\) is given by

\[
\Pr(\omega_1, \omega_2) = \Pr_1(\omega_1) \Pr_2(\omega_2), \quad \omega_1 \in \Omega_1, \omega_2 \in \Omega_2.
\]

There is an obvious generalization for \(n\) discrete probability spaces. In particular, for any discrete probability space \((\Omega, \mathcal{F}, \Pr)\) and any integer \(n \geq 1\), we define the product space \((\Omega^n, \mathcal{F}), \Pr)\), with

\[
\Pr(\omega_1, \ldots, \omega_n) = \Pr(\omega_1) \cdots \Pr(\omega_n), \quad \omega_i \in \Omega, i = 1, \ldots, n.
\]

The fact that the probability measure on the product space is defined as a product of the probability measures of its components captures the independence of the trials.

**Remark:** The product of two probability spaces \((\Omega_1, \mathcal{F}_1, \Pr_1)\) and \((\Omega_2, \mathcal{F}_2, \Pr_2)\) can also be defined, but \(\mathcal{F}_1 \times \mathcal{F}_2\) is not a \(\sigma\)-algebra in general, so some serious work needs to be done.

The notion of independence also applies to random variables. Given two random variables \(X\) and \(Y\) on the same (discrete) probability space, it is useful to consider their **joint distribution** (really joint mass function) \(f_{X,Y}\) given by

\[
f_{X,Y}(a, b) = \Pr(X = a \text{ and } Y = b) = \Pr(\{\omega \in \Omega \mid (X(\omega) = a) \wedge (Y(\omega) = b)\}),
\]
for any two reals \(a, b \in \mathbb{R}\).

**Definition 6.7.** Two random variables \(X\) and \(Y\) defined on the same discrete probability space are **independent** if

\[
\Pr(X = a \text{ and } Y = b) = \Pr(X = a) \Pr(Y = b), \quad \text{for all } a, b \in \mathbb{R}.
\]

**Remark:** If \(X\) and \(Y\) are two continuous random variables, we say that \(X\) and \(Y\) are **independent** if

\[
\Pr(X \leq a \text{ and } Y \leq b) = \Pr(X \leq a) \Pr(Y \leq b), \quad \text{for all } a, b \in \mathbb{R}.
\]
It is easy to verify that if $X$ and $Y$ are discrete random variables, then the above condition is equivalent to the condition of Definition 6.7.

Example 6.15. If we consider the probability space of Example 6.2 (rolling two dice), then we can define two random variables $S_1$ and $S_2$, where $S_1$ is the value on the first dice and $S_2$ is the value on the second dice. Then, the total of the two values is the random variable $S = S_1 + S_2$ of Example 6.6. Since

$$\Pr(S_1 = a \text{ and } S_2 = b) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \Pr(S_1 = a)\Pr(S_2 = b),$$

the random variables $S_1$ and $S_2$ are independent.

Example 6.16. Suppose we flip a biased coin (with probability $p$ of success) $n$ times. Let $X$ be the number of heads observed and let $Y$ be the number of tails observed. The variables $X$ and $Y$ are not independent. For example

$$\Pr(X = 1 \text{ and } Y = 1) = 1,$$

yet

$$\Pr(X = 1)\Pr(Y = 1) = p(1 - p).$$

It is remarkable that if we assume the Poisson paradigm, that $X$ and $Y$ have Poisson distribution with parameter $\lambda$, then it can be shown that $X$ and $Y$ are independent.

The following characterization of independence for two random variables is left as an exercise.

**Proposition 6.4.** If $X$ and $Y$ are two random variables on a discrete probability space $(\Omega, \Pr)$ and if $f_{X,Y}$ is the joint distribution (mass function) of $X$ and $Y$, $f_X$ is the distribution (mass function) of $X$ and $f_Y$ is the distribution (mass function) of $Y$, then $X$ and $Y$ are independent iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$  

**Remark:** To deal with the continuous case, it is useful to consider the joint distribution $F_{X,Y}$ of $X$ and $Y$ given by

$$F_{X,Y}(a,b) = \Pr(X \leq a \text{ and } Y \leq b) = \Pr(\{\omega \in \Omega \mid (X(\omega) \leq a) \land (Y(\omega) \leq b)\}),$$

for any two reals $a, b \in \mathbb{R}$. We say that $X$ and $Y$ are jointly continuous with joint density function $f_{X,Y}$ if

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) du dv, \quad \text{for all } x, y \in \mathbb{R}$$

for some nonnegative integrable function $f_{X,Y}$. The marginal density functions $f_X$ of $X$ and $f_Y$ of $Y$ are defined by
6.4 Expectation of a Random Variable

\[ f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx. \]

They correspond to the marginal distribution functions \( F_X \) of \( X \) and \( F_Y \) of \( Y \) given by

\[ F_X(x) = \Pr(X \leq x) = F_{X,Y}(x,\infty), \quad F_Y(y) = \Pr(Y \leq y) = F_{X,Y}(\infty,y). \]

Then, it can be shown that \( X \) and \( Y \) are independent iff

\[ F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}, \]

which, for continuous variables, is equivalent to

\[ f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y \in \mathbb{R}. \]

We now turn to one of the most important concepts about random variables, their mean (or expectation).

6.4 Expectation of a Random Variable

In order to understand the behavior of a random variable, we may want to look at its “average” value. But the notion of average is ambiguous, as there are different kinds of averages that we might want to consider. Among these, we have

1. the mean: the sum of the values divided by the number of values.
2. the median: the middle value (numerically).
3. the mode: the value that occurs most often.

For example, the mean of the sequence \((3, 1, 4, 1, 5)\) is 2.8; the median is 3, and the mode is 1.

Given a random variable \( X \), if we consider a sequence of values \( X(\omega_1), X(\omega_2), \ldots, X(\omega_n) \), each value \( X(\omega_j) = a_j \) has a certain probability \( \Pr(X = a_j) \) of occurring which may differ depending on \( j \), so the usual mean

\[ \frac{X(\omega_1) + X(\omega_2) + \cdots + X(\omega_n)}{n} = \frac{a_1 + \cdots + a_n}{n} \]

may not capture well the “average” of the random variable \( X \). A better solution is to use a weighted average, where the weights are probabilities. If we write \( a_j = X(\omega_j) \), we can define the mean of \( X \) as the quantity

\[ a_1\Pr(X = a_1) + a_2\Pr(X = a_2) + \cdots + a_n\Pr(X = a_n). \]
Definition 6.8. Given a discrete probability space \((\Omega, \Pr)\), for any random variable \(X\), the mean value or expected value or expectation\(^1\) of \(X\) is the number \(E(X)\) defined as

\[ E(X) = \sum_{x \in X(\Omega)} x \cdot \Pr(X = x) = \sum_{x|f(x)>0} xf(x), \]

where \(X(\Omega)\) denotes the image of the function \(X\) and where \(f\) is the probability mass function of \(X\). Because \(\Omega\) is finite, we can also write

\[ E(X) = \sum_{\omega \in \Omega} X(\omega)\Pr(\omega). \]

In this setting, the median of \(X\) is defined as the set of elements \(x \in \Omega\) such that

\[ \Pr(X \leq x) \geq \frac{1}{2} \quad \text{and} \quad \Pr(X \geq x) \geq \frac{1}{2}. \]

Remark: If \(\Omega\) is countably infinite, then the expectation \(E(X)\), if it exists, is given by

\[ E(X) = \sum_{x|f(x)>0} xf(x), \]

provided that the above sum converges absolutely (that is, the partial sums of absolute values converge). If we have a probability space \((X, \mathcal{F}, \Pr)\) with \(\Omega\) uncountable and if \(X\) is absolutely continuous so that it has a density function \(f\), then the expectation of \(X\) is given by the integral

\[ E(X) = \int_{-\infty}^{+\infty} xf(x)dx. \]

It is even possible to define the expectation of a random variable that is not necessarily absolutely continuous using its cumulative density function \(F\) as

\[ E(X) = \int_{-\infty}^{+\infty} xdF(x), \]

where the above integral is the Lebesgue–Stieltjes integral, but this is way beyond the scope of this book.

Observe that if \(X\) is a constant random variable (that is, \(X(\omega) = c\) for all \(\omega \in \Omega\) for some constant \(c\), then

\[ E(X) = \sum_{\omega \in \Omega} X(\omega)\Pr(\omega) = c \sum_{\omega \in \Omega} \Pr(\omega) = c\Pr(\Omega) = c, \]

since \(\Pr(\Omega) = 1\). The mean of a constant random variable is itself (as it should!).

\(^1\) It is amusing that in French, the word for expectation is espérance mathématique. There is hope for mathematics!
Example 6.17. Consider the sum $S$ of the values on the dice from Example 6.6. The expectation of $S$ is

$$E(S) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + \cdots + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + \cdots + 12 \cdot \frac{1}{36} = 7.$$ 

Example 6.18. Suppose we flip a biased coin once (with probability $p$ of landing tails). If $X$ is the random variable given by $X(H) = 1$ and $X(T) = 0$, the expectation of $X$ is

$$E(X) = 1 \cdot \Pr(X = 1) + 0 \cdot \Pr(X = 0) = 1 \cdot P + 0 \cdot (1 - p) = p.$$ 

Example 6.19. Consider the binomial distribution of Example 6.8, where the random variable $X$ counts the number of tails (success) in a sequence of $n$ trials. Let us compute $E(X)$. Since the mass function is given by

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

we have

$$E(X) = \sum_{i=0}^{n} if(i) = \sum_{i=0}^{n} i \binom{n}{i} p^i (1 - p)^{n-i}.$$ 

We use a trick from analysis to compute this sum. Recall from the binomial theorem that

$$(1+x)^n = \sum_{i=0}^{n} \binom{n}{i} x^i.$$ 

If we take derivatives on both sides, we get

$$n(1+x)^{n-1} = \sum_{i=0}^{n} i \binom{n}{i} x^{i-1},$$

and by multiplying both sides by $x$,

$$nx(1+x)^{n-1} = \sum_{i=0}^{n} i \binom{n}{i} x^i.$$ 

Now, if we set $x = p/q$, since $p + q = 1$, we get

$$\sum_{i=0}^{n} i \binom{n}{i} p^i (1 - p)^{n-i} = np,$$

and so

$$E(X) = np.$$ 

A crucial property of expectation that often allows simplifications in computing the expectation of a random variable is its linearity.
Proposition 6.5. (Linearity of Expectation) Given two random variables on a discrete probability space, for any real number \( \lambda \), we have

\[
\begin{align*}
E(X+Y) &= E(X) + E(Y) \\
E(\lambda X) &= \lambda E(X).
\end{align*}
\]

Proof. We have

\[
E(X+Y) = \sum_z z \cdot \Pr(X+Y = z)
\]

\[
= \sum_x \sum_y (x+y) \cdot \Pr(X=x \text{ and } Y=y)
\]

\[
= \sum_x \sum_y x \cdot \Pr(X=x \text{ and } Y=y) + \sum_y \sum_x y \cdot \Pr(X=x \text{ and } Y=y)
\]

\[
= \sum_x \sum_y \Pr(X=x \text{ and } Y=y) + \sum_y \sum_x \Pr(X=x \text{ and } Y=y).
\]

Now, the events \( A_x = \{ x \mid X = x \} \) form a partition of \( \Omega \), which implies that

\[
\sum_y \Pr(X=x \text{ and } Y=y) = \Pr(X=x).
\]

Similarly, the events \( B_y = \{ y \mid Y = y \} \) form a partition of \( \Omega \), which implies that

\[
\sum_x \Pr(X=x \text{ and } Y=y) = \Pr(Y=y).
\]

By substitution, we obtain

\[
E(X+Y) = \sum_x x \cdot \Pr(X=x) + \sum_y y \cdot \Pr(Y=y),
\]

proving that \( E(X+Y) = E(X) + E(Y) \). When \( \Omega \) is countably infinite, we can permute the indices \( x \) and \( y \) due to absolute convergence.

For the second equation, if \( \lambda \neq 0 \), we have

\[
E(\lambda X) = \sum_x x \cdot \Pr(\lambda X = x)
\]

\[
= \lambda \sum_x \frac{x}{\lambda} \cdot \Pr(X=x/\lambda)
\]

\[
= \lambda \sum_y y \cdot \Pr(X=y)
\]

\[
= \lambda E(X).
\]

as claimed. If \( \lambda = 0 \), the equation is trivial. \( \square \)
By a trivial induction, we obtain that for any finite number of random variables $X_1, \ldots, X_n$, we have

$$E\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} E(X_i).$$

It is also important to realize that the above equation holds even if the $X_i$ are not independent.

Here is an example showing how the linearity of expectation can simplify calculations. Let us go back to Example 6.19. Define $n$ random variables $X_1, \ldots, X_n$ such that $X_i(\omega) = 1$ iff the $i$th flip yields heads, otherwise $X_i(\omega) = 0$. Clearly, the number $X$ of heads in the sequence is

$$X = X_1 + \cdots + X_n.$$ 

However, we saw in Example 6.18 that $E(X_i) = p$, and since

$$E(X) = E(X_1) + \cdots + E(X_n),$$

we get

$$E(X) = np.$$ 

The above example suggests the definition of indicator function, which turns out to be quite handy.

**Definition 6.9.** Given a discrete probability space with sample space $\Omega$, for any event $A$, the indicator function (or indicator variable) of $A$ is the random variable $I_A$ defined such that

$$I_A(\omega) = \begin{cases} 
1 & \text{if } \omega \in A \\
0 & \text{if } \omega \notin A.
\end{cases}$$

The main property of the indicator function $I_A$ is that its expectation is equal to the probability $Pr(A)$ of the event $A$. Indeed,

$$E(I_A) = \sum_{\omega \in \Omega} I_A(\omega)Pr(\omega) = \sum_{\omega \in A} Pr(\omega) = Pr(A).$$

This fact with the linearity of expectation is often used to compute the expectation of a random variable, by expressing it as a sum of indicator variables. We will see how this method is used to compute the expectation of the number of comparisons in quicksort. But first, we use this method to find the expected number of fixed points of a random permutation.

**Example 6.20.** For any integer $n \geq 1$, let $\Omega$ be the set of all $n!$ permutations of $\{1, \ldots, n\}$, and give $\Omega$ the uniform probability measure; that is, for every permutation $\pi$, let
We say that these are random permutations. A fixed point of a permutation \( \pi \) is any integer \( k \) such that \( \pi(k) = k \). Let \( X \) be the random variable such that \( X(\pi) \) is the number of fixed points of the permutation \( \pi \). Let us find the expectation of \( X \). To do this, for every \( k \), let \( X_k \) be the random variable defined so that \( X_k(\pi) = 1 \) iff \( \pi(k) = k \), and 0 otherwise. Clearly,

\[
X = X_1 + \cdots + X_n,
\]

and since

\[
E(X) = E(X_1) + \cdots + E(X_n),
\]

we just have to compute \( E(X_k) \). But, \( X_k \) is an indicator variable, so

\[
E(X_k) = Pr(X_k = 1) = \frac{1}{n}.
\]

Now, there are \((n - 1)!\) permutations that leave \( k \) fixed, so \( Pr(X = 1) = \frac{1}{n} \). Therefore,

\[
E(X) = E(X_1) + \cdots + E(X_n) = n \cdot \frac{1}{n} = 1.
\]

On average, a random permutation has one fixed point.

If \( X \) is a random variable on a discrete probability space \( \Omega \) (possibly countably infinite), for any function \( g : \mathbb{R} \to \mathbb{R} \), the composition \( g \circ X \) is a random variable defined by

\[
(g \circ X)(\omega) = g(X(\omega)), \quad \omega \in \Omega.
\]

This random variable is usually denoted by \( g(X) \).

Given two random variables \( X \) and \( Y \), if \( \varphi \) and \( \psi \) are two functions, we leave it as an exercise to prove that if \( X \) and \( Y \) are independent, then so are \( \varphi(X) \) and \( \psi(Y) \).

Although computing its mass function in terms of the mass function \( f \) of \( X \) can be very difficult, there is a nice way to compute its expectation.

**Proposition 6.6.** If \( X \) is a random variable on a discrete probability space \( \Omega \), for any function \( g : \mathbb{R} \to \mathbb{R} \), the expectation \( E(g(X)) \) of \( g(X) \) (if it exists) is given by

\[
E(g(X)) = \sum_x g(x) f(x),
\]

where \( f \) is the mass function of \( X \).

**Proof.** We have
6.4 Expectation of a Random Variable

\[ E(g(X)) = \sum_y y \cdot Pr(g \circ X = y) \]
\[ = \sum_y y \cdot Pr(\{ \omega \in \Omega \mid g(X(\omega)) = y \}) \]
\[ = \sum_y \sum_x y \cdot Pr(\{ \omega \in \Omega, g(x) = y, X(\omega) = x \}) \]
\[ = \sum_y \sum_{x,g(x)=y} y \cdot Pr(\{ \omega \in \Omega, X(\omega) = x \}) \]
\[ = \sum_y g(x) \cdot Pr(X = x) \]
\[ = \sum_x g(x) f(x), \]

as claimed.

The cases \( g(X) = X^k \), \( g(X) = z^X \), and \( g(X) = e^{tX} \) (for some given reals \( z \) and \( t \)) are of particular interest.

**Example 6.21.** Consider the random variable \( X \) of Example 6.19 counting the number of heads in a sequence of coin flips of length \( n \), but this time, let us try to compute \( E(X^k) \), for \( k \geq 2 \). We have

\[ E(X^k) = \sum_{i=0}^{\infty} i^k f(i) \]
\[ = \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \]
\[ = \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}. \]

Recall that
\[ i \binom{n}{i} = n \binom{n-1}{i-1}. \]

Using this, we get

\[ E(X^k) = \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \]
\[ = np \sum_{i=1}^{n} j^{i-1} \binom{n-1}{i-1} p^j (1-p)^{n-1-j} \]
\[ = np \sum_{j=0}^{n-1} (j+1)^{i-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \]
\[ = np E((Y + 1)^{i-1}), \]
where \( Y \) is a random variable with binomial distribution on sequences of length \( n - 1 \) and with the same probability \( p \) of success. Thus, we obtain an inductive method to compute \( \mathbb{E}(X^k) \). For \( k = 2 \), we get

\[
\mathbb{E}(X^2) = np\mathbb{E}(Y + 1) = np((n - 1)p + 1).
\]

If \( X \) only takes nonnegative integer values, then the following result may be useful for computing \( \mathbb{E}(X) \).

**Proposition 6.7.** If \( X \) is a random variable that takes on only nonnegative integers, then its expectation \( \mathbb{E}(X) \) (if it exists) is given by

\[
\mathbb{E}(X) = \sum_{i=1}^{\infty} \text{Pr}(X \geq i).
\]

**Proof.** We have

\[
\sum_{i=1}^{\infty} \text{Pr}(X \geq i) = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \text{Pr}(X = j)
\]

\[
= \sum_{j=1}^{\infty} \sum_{i=1}^{j} \text{Pr}(X = j)
\]

\[
= \sum_{j=1}^{\infty} j \text{Pr}(X = j)
\]

\[
= \mathbb{E}(X).
\]

The interchange of infinite summations is legitimate because all the terms are nonnegative.

Here is an application of Proposition 6.7.

**Example 6.22.** In Example 6.9, we consider finite sequences of flips of a biased coin, and the random variable of interest is the first occurrence of tails (success). The distribution of this random variable is the geometric distribution,

\[
f(n) = (1 - p)^{n-1} p, \quad n \geq 1.
\]

To compute its expectation, let us use Proposition 6.7. We have
6.4 Expectation of a Random Variable

\[
\Pr(X \geq i) = \sum_{i=1}^{\infty} (1 - p)^{i-1} p
\]
\[
= p(1 - p)^{i-1} \sum_{j=0}^{\infty} (1 - p)^j
\]
\[
= p(1 - p)^{i-1} \frac{1}{1 - (1 - p)}
\]
\[
= (1 - p)^{i-1}.
\]

Then, we have

\[
E(X) = \sum_{i=1}^{\infty} \Pr(X \geq i)
\]
\[
= \sum_{i=1}^{\infty} (1 - p)^{i-1}.
\]
\[
= \frac{1}{1 - (1 - p)} = \frac{1}{p}.
\]

Therefore,

\[
E(X) = \frac{1}{p},
\]

which means that on the average, it takes \(1/p\) flips until heads turns up.

Let us now compute \(E(X^2)\). We have

\[
E(X^2) = \sum_{i=1}^{\infty} i^2 (1 - p)^{i-1} p
\]
\[
= \sum_{i=1}^{\infty} (i - 1 + 1)^2 (1 - p)^{i-1} p
\]
\[
= \sum_{i=1}^{\infty} (i - 1)^2 (1 - p)^{i-1} p + \sum_{i=1}^{\infty} 2(i - 1)(1 - p)^{i-1} p + \sum_{i=1}^{\infty} (1 - p)^{i-1} p
\]
\[
= \sum_{j=0}^{\infty} j^2 (1 - p)^j p + 2 \sum_{j=1}^{\infty} j(1 - p)^j p + 1 \quad \text{(let } j = i - 1) \]
\[
= (1 - p)E(X^2) + 2(1 - p)E(X) + 1.
\]

Since \(E(X) = 1/p\), we obtain

\[
pE(X^2) = \frac{2(1 - p)}{p} + 1
\]
\[
= \frac{2 - p}{p},
\]

so
\[ E(X^2) = \frac{2-p}{p^2}. \]

By the way, the trick of writing \( i = i - 1 + 1 \) can be used to compute \( E(X) \). Try to recompute \( E(X) \) this way.

**Example 6.23.** Let us compute the expectation of the number \( X \) of comparisons needed when running the randomized version of quicksort presented in Example 6.7. Recall that the input is a sequence \( S = (x_1, \ldots, x_n) \) of distinct elements, and that \( (y_1, \ldots, y_n) \) has the same elements sorted in increasing order. In order to compute \( E(X) \), we decompose \( X \) as a sum of indicator variables \( X_{i \rightarrow j} \), with \( X_{i \rightarrow j} = 1 \) iff \( y_i \) and \( y_j \) are ever compared, and \( X_{i \rightarrow j} = 0 \) otherwise. Then, it is clear that

\[ X = \sum_{j=2}^{n} \sum_{i=1}^{j-1} X_{i \rightarrow j}, \]

and

\[ E(X) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} E(X_{i \rightarrow j}). \]

Furthermore, since \( X_{i \rightarrow j} \) is an indicator variable, we have

\[ E(X_{i \rightarrow j}) = \Pr(y_i \text{ and } y_j \text{ are ever compared}). \]

The crucial observation is that \( y_i \) and \( y_j \) are ever compared iff either \( y_i \) or \( y_j \) is chosen as the pivot when \( \{y_i, y_{i+1}, \ldots, y_j\} \) is a subset of the set of elements of the (left or right) sublist considered for the choice of a pivot.

This is because if the next pivot \( y \) is larger than \( y_j \), then all the elements in \( \{y_i, y_{i+1}, \ldots, y_j\} \) are placed in the list to the left of \( y \), and if \( y \) is smaller than \( y_i \), then all the elements in \( \{y_i, y_{i+1}, \ldots, y_j\} \) are placed in the list to the right of \( y \). Consequently, if \( y_i \) and \( y_j \) are ever compared, some pivot \( y \) must belong to \( \{y_i, y_{i+1}, \ldots, y_j\} \), and every \( y_k \neq y \) in the list will be compared with \( y \). But, if the pivot \( y \) is distinct from \( y_i \) and \( y_j \), then \( y_i \) is placed in the left sublist and \( y_j \) in the right sublist, so \( y_i \) and \( y_j \) will never be compared.

It remains to compute the probability that the next pivot chosen in the sublist \( Y_{i,j} = \{y_i, y_{i+1}, \ldots, y_j\} \) is \( y_i \) (or that the next pivot chosen is \( y_j \), but the two probabilities are equal). Since the pivot is one of the values in \( \{y_i, y_{i+1}, \ldots, y_j\} \) and since each of these is equally likely to be chosen (by hypothesis), we have

\[ \Pr(y_i \text{ is chosen as the next pivot in } Y_{i,j}) = \frac{1}{j-i+1}. \]

Consequently, since \( y_i \) and \( y_j \) are ever compared iff either \( y_i \) is chosen as a pivot or \( y_j \) is chosen as a pivot, and since these two events are mutually exclusive, we have

\[ E(X_{i \rightarrow j}) = \Pr(y_i \text{ and } y_j \text{ are ever compared}) = \frac{2}{j-i+1}. \]
It follows that

\[ E(X) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} E(X_{i,j}) \]

\[ = 2 \sum_{j=2}^{n} \sum_{k=1}^{j} \frac{1}{k} \quad \text{(set } k = j - i + 1) \]

\[ = 2 \sum_{k=2}^{n} \sum_{j=k}^{n} \frac{1}{k} \]

\[ = 2 \sum_{k=2}^{n} \frac{n - k + 1}{k} \]

\[ = 2(n + 1) \sum_{k=1}^{n} \frac{1}{k} - 4n. \]

At this stage, we use the result of Problem 5.32. Indeed,

\[ \sum_{k=1}^{n} \frac{1}{k} = H_n \]

is a harmonic number, and it is shown that

\[ \ln(n) + \frac{1}{2n} \leq H_n \leq \ln(n + 1) + \frac{1}{2n}. \]

Therefore, \( H_n = \ln(n) + \Theta(1) \), which shows that

\[ E(X) = 2n \log n + \Theta(n). \]

Therefore, the expected number of comparisons made by the randomized version of quicksort is \( 2n \log n + \Omega(n) \).

**Example 6.24.** If \( X \) is a random variable with Poisson distribution with parameter \( \lambda \) (see Example 6.10), let us show that its expectation is

\[ E(X) = \lambda. \]

Recall that a Poisson distribution is given by

\[ f(i) = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i \in \mathbb{N}, \]

so we have
\[
\begin{align*}
E(X) &= \sum_{i=0}^{\infty} i e^{-\lambda} \frac{\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad \text{(let } j = i - 1) \\
&= \lambda e^{-\lambda} e^\lambda = \lambda,
\end{align*}
\]

as claimed. This is consistent with the fact that the expectation of a random variable with a binomial distribution is \(np\), under the Poisson approximation where \(\lambda = np\). We leave it as an exercise to prove that
\[
E(X^2) = \lambda(\lambda + 1).
\]

Although in general \(E(XY) \neq E(X)E(Y)\), this is true for independent random variables.

**Proposition 6.8.** If two random variables \(X\) and \(Y\) on the same discrete probability space are independent, then
\[
E(XY) = E(X)E(Y).
\]

**Proof.** We have
\[
\begin{align*}
E(XY) &= \sum_{\omega \in \Omega} X(\omega)Y(\omega) \Pr(\omega) \\
&= \sum_{x} \sum_{y} xy \cdot \Pr(X = x \text{ and } Y = y) \\
&= \sum_{x} \sum_{y} xy \cdot \Pr(X = x) \Pr(Y = y) \\
&= \left( \sum_{x} x \cdot \Pr(X = x) \right) \left( \sum_{y} y \cdot \Pr(Y = y) \right) \\
&= E(X)E(Y),
\end{align*}
\]

as claimed. Note that the independence of \(X\) and \(Y\) was used in going from line 2 to line 3. \(\square\)

In Example 6.15 (rolling two dice), we defined the random variables \(S_1\) and \(S_2\), where \(S_1\) is the value on the first dice and \(S_2\) is the value on the second dice. We also showed that \(S_1\) and \(S_2\) are independent. If we consider the random variable \(P = S_1S_2\), then we have
\[
E(P) = E(S_1)E(S_2) = \frac{7}{2} \cdot \frac{7}{2} = \frac{49}{4},
\]
since \(E(S_1) = E(S_2) = 7/2\), as we easily determine since all probabilities are equal to 1/6. On the other hand, \(S\) and \(P\) are not independent (check it).