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 phenomema 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, Aleksandr Lyapunov, Andrei Markov, Emile Borel, and Paul Lévy should also be added to the list.

Fig. 6.1 Pierre de Fermat (1601–1665) (left), Blaise Pascal (1623–1662) (middle left), Pierre–Simon Laplace (1749–1827) (middle right), Andrey Nikolaevich Kolmogorov (1903–1987) (right)

Before Kolmogorov, probability theory was a subject that still lacked precise definitions. In1933, 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

$$\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\},$$

where H stands for heads and T stands for tails.

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

$$\Omega = \{\mathrm{HHHHH}, \mathrm{THHHH}, \mathrm{HTHHH}, \mathrm{TTHHH}, \ldots, \mathrm{TTTTT}\}.$$

Example 6.2. If the experiment consists in rolling a pair of dice, then the sample space Ω consists of the 36 pairs in the set

$$\Omega = D \times D$$

with

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

where the integer $i \in D$ corresponds to the number (indicated by dots) on the face of the dice facing up, as shown in Figure 6.2. 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 Ω be the sample space of all possible hands. This time it is not possible to enumerate the sample space explicitly. Indeed, there are

Fig. 6.2 Two dice

$$\binom{52}{13} = \frac{52!}{13! \cdot 39!} = \frac{52 \cdot 51 \cdot 50 \cdots 40}{13 \cdot 12 \cdot \cdots \cdot 2 \cdot 1} = 635,013,559,600$$

different hands, a huge number.

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

$$A = \{\text{HTTTT}, \text{THTTT}, \text{TTHTT}, \text{TTTHT}, \text{TTTTH}\}.$$

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 outcome ω of a sample space Ω is assigned some probability $\mathsf{Pr}(\omega)$. Intuitively, $\mathsf{Pr}(\omega)$ is the probabilty that the outcome ω may occur. It is convenient to normalize probabilites, so we require that

$$0 \leq \mathsf{Pr}(\omega) \leq 1.$$

If Ω is finite, we also require that

$$\sum_{\omega \in \Omega} \mathsf{Pr}(\omega) = 1.$$

The function Pr is often called a *probability distribution* on Ω. Indeed, it distributes the probability of 1 among the outomes ω.

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

For example, if we assume that our coins are "fair," then when we flip a coin five times, since each outcome in Ω is equally likely, the probability of each outcome $\omega \in \Omega$ is

$$\mathsf{Pr}(\omega) = \frac{1}{32}.$$

If we assume that our dice are "fair," namely that each of the six possibilities for a particular dice has probability $1/6$, 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

$$\Pr_1(1) = \Pr_1(6) = \frac{1}{4}$$
$$\Pr_1(2) = \Pr_1(3) = \Pr_1(4) = \Pr_1(5) = \frac{1}{8}.$$

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 Ω. 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 Ω of *outcomes* or *elementary events* $\omega \in \Omega$, together with a function $\Pr\colon \Omega \to \mathbb{R}$, called *probability measure* (or *probability distribution*) satisfying the following properties:

$$0 \le \Pr(\omega) \le 1 \quad \text{for all } \omega \in \Omega.$$
$$\sum_{\omega\in\Omega} \Pr(\omega) = 1.$$

An *event* is any subset A of Ω. The probability of an event A is defined as

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

Definition 6.1 immediately implies that

$$\Pr(\emptyset) = 0$$
$$\Pr(\Omega) = 1.$$

For another example, if we consider the event

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

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

$$\mathsf{Pr}(A) = \frac{5}{32}.$$

If we use the probability distribution Pr on the sample space Ω 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

$$\mathsf{Pr}(B) = 6 \cdot \frac{1}{36} = \frac{1}{6}.$$

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

$$\mathsf{Pr}_{11}(B) = \frac{1}{16} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{16} = \frac{3}{16} > \frac{1}{16}.$$

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 Ω 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 $(\Omega, \mathscr{F}, \mathsf{Pr})$ consisting of:

1. A nonempty countably infinite set Ω of *outcomes* or *elementary events*.
2. The set $\mathscr{F}$ of all subsets of Ω, called the set of *events*.
3. A function $\mathsf{Pr}\colon \mathscr{F} \to \mathbb{R}$, called *probability measure* (or *probability distribution*) satisfying the following properties:

   a. (positivity)
   $$0 \leq \mathsf{Pr}(A) \leq 1 \quad \text{for all } A \in \mathscr{F}.$$

   b. (normalization)
   $$\mathsf{Pr}(\Omega) = 1.$$

   c. (additivity and continuity)
   For any sequence of pairwise disjoint events $E_1, E_2, \ldots, E_i, \ldots$ in $\mathscr{F}$ (which means that $E_i \cap E_j = \emptyset$ for all $i \neq j$), we have
   $$\mathsf{Pr}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathsf{Pr}(E_i).$$

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

$$\mathsf{Pr}(\emptyset) = 0.$$

The notion of a discrete probability space is sufficient to deal with most problems that a computer scientist or an engineer will ever encounter. However, there are certain problems for which it is necessary to assume that the family $\mathscr{F}$ of events is a proper subset of the power set of Ω. In this case, $\mathscr{F}$ is called the family of *measurable* events, and $\mathscr{F}$ has certain closure properties that make it a σ*-algebra* (also called a σ*-field*). Some problems even require Ω to be uncountably infinite. In this case, we drop the word *discrete* from discrete probability space.

Remark: A σ*-algebra* is a nonempty family $\mathscr{F}$ of subsets of Ω satisfying the following properties:

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

Note that every σ-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 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 $\mathscr{F} = 2^{\Omega}$, and for notational simplicity we omit $\mathscr{F}$ (that is, we write (Ω, Pr) instead of $(\Omega, \mathscr{F}, \mathsf{Pr})$).

Because events are subsets of the sample space Ω, they can be combined using the set operations, union, intersection, and complementation. If the sample space Ω is finite, the definition for the probability $\mathsf{Pr}(A)$ 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

$$\mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B).$$

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

$$\mathsf{Pr}(A_1 \cup \cdots \cup A_n) = \mathsf{Pr}(A_1) + \cdots + \mathsf{Pr}(A_n).$$

It is natural to ask whether the probabilities $\mathsf{Pr}(A \cup B)$, $\mathsf{Pr}(A \cap B)$ and $\mathsf{Pr}(\overline{A})$ can be expressed in terms of $\mathsf{Pr}(A)$ and $\mathsf{Pr}(B)$, 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* (Ω, Pr)*, for any two events* $A, B \subseteq \Omega$*, we have*

$$\mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(A \cap B)$$
$$\mathsf{Pr}(\overline{A}) = 1 - \mathsf{Pr}(A).$$

Furthermore, if $A \subseteq B$*, then* $\mathsf{Pr}(A) \leq \mathsf{Pr}(B)$.

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

$$\begin{aligned} \mathsf{Pr}(A \cup B) &= \mathsf{Pr}(A \cap B) + \mathsf{Pr}(A - B) + \mathsf{Pr}(B - A) \\ \mathsf{Pr}(A) &= \mathsf{Pr}(A \cap B) + \mathsf{Pr}(A - B) \\ \mathsf{Pr}(B) &= \mathsf{Pr}(A \cap B) + \mathsf{Pr}(B - A), \end{aligned}$$

and from these we obtain

$$\mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(A \cap B).$$

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

$$1 = \mathsf{Pr}(\Omega) = \mathsf{Pr}(A) + \mathsf{Pr}(\overline{A}).$$

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

$$\mathsf{Pr}(B) = \mathsf{Pr}(A) + \mathsf{Pr}(B - A).$$

Since probabilities are nonegative, the above implies that $\mathsf{Pr}(A) \leq \mathsf{Pr}(B)$. □

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

$$\mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(A \cap B)$$

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).

The following proposition expresses a certain form of continuity of the function Pr.

Proposition 6.2. *Given any probability space $(\Omega, \mathscr{F}, \mathsf{Pr})$ (discrete or not), for any sequence of events $(A_i)_{i \geq 1}$, if $A_i \subseteq A_{i+1}$ for all $i \geq 1$, then*

$$\mathsf{Pr}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \mapsto \infty} \mathsf{Pr}(A_n).$$

Proof. The trick is to express $\bigcup_{i=1}^{\infty} A_i$ as a union of pairwise disjoint events. Indeed, we have

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 - A_1) \cup (A_3 - A_2) \cup \cdots \cup (A_{i+1} - A_i) \cup \cdots,$$

so by property (c) of a probability measure

$$\begin{aligned}\mathsf{Pr}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mathsf{Pr}\left(A_1 \cup \bigcup_{i=1}^{\infty}(A_{i+1} - A_i)\right) \\ &= \mathsf{Pr}(A_1) + \sum_{i=1}^{\infty} \mathsf{Pr}(A_{i+1} - A_i) \\ &= \mathsf{Pr}(A_1) + \lim_{n \mapsto \infty} \sum_{i=1}^{n-1} \mathsf{Pr}(A_{i+1} - A_i) \\ &= \lim_{n \mapsto \infty} \mathsf{Pr}(A_n),\end{aligned}$$

as claimed.

We leave it as an exercise to prove that if $A_{i+1} \subseteq A_i$ for all $i \geq 1$, then

$$\mathsf{Pr}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n \mapsto \infty} \mathsf{Pr}(A_n).$$

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

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

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

Two events are *dependent* if they are not independent.

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

$$A = \{\mathrm{HH}w \mid w \in \{\mathrm{H},\mathrm{T}\}^3\} \cup \{\mathrm{HT}w \mid w \in \{\mathrm{H},\mathrm{T}\}^3\},$$

the event in which the first flip is H, and

$$B = \{\mathrm{HH}w \mid w \in \{\mathrm{H},\mathrm{T}\}^3\} \cup \{\mathrm{TH}w \mid w \in \{\mathrm{H},\mathrm{T}\}^3\},$$

the event in which the second flip is H. Since A and B contain 16 outcomes, we have

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

The intersection of A and B is

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

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

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

Since

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

and

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

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

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

and

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

we have

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

and since

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

we have

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

It follows that

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

but

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

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 365 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 n distinct birthdays in $\binom{365}{n}$ 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 probabilty 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 $x \leq \mathrm{e}^{x-1}$ for all $x \in \mathbb{R}$, so $1 - x \leq \mathrm{e}^{-x}$ for all $x \in \mathbb{R}$, and we can bound q as follows:

$$\begin{aligned} q &= \prod_{i=1}^{n-1} \left(1 - \frac{i}{365}\right) \\ q &\leq \prod_{i=1}^{n-1} \mathrm{e}^{-i/365} \\ &= \mathrm{e}^{-\sum_{i=1}^{n-1} \frac{i}{365}} \\ &\mathrm{e}^{-\frac{n(n-1)}{2 \cdot 365}}. \end{aligned}$$

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

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

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 $p = 50.7\%$. 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 (Ω, Pr), we are more interested in the behavior of functions $X\colon \Omega \to \mathbb{R}$ defined on the sample space Ω 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 Ω, thus an event, so we may consider the probability $\mathsf{Pr}(X^{-1}(a))$, denoted (somewhat improperly) by

$$\mathsf{Pr}(X = a).$$

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 ω is a random variable. We determine that

$$\mathsf{Pr}(X=0) = \frac{1}{32} \qquad \mathsf{Pr}(X=1) = \frac{5}{32} \qquad \mathsf{Pr}(X=2) = \frac{10}{32}$$
$$\mathsf{Pr}(X=3) = \frac{10}{32} \qquad \mathsf{Pr}(X=4) = \frac{5}{32} \qquad \mathsf{Pr}(X=5) = \frac{1}{32}.$$

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

$$\mathsf{Pr}(Y=0) = \frac{1}{32}$$
$$\mathsf{Pr}(Y=1) = \frac{31}{32}.$$

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 $\mathsf{Pr}(S = s)$ for $s = 2, \ldots, 12$, we find the following table:

s	2	3	4	5	6	7	8	9	10	11	12
$\mathsf{Pr}(S=s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Here is a "real" example from computer science.

Example 6.7. Our goal is to sort of a sequence $S = (x_1, \ldots, x_n)$ of n distinct real numbers 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, x, 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

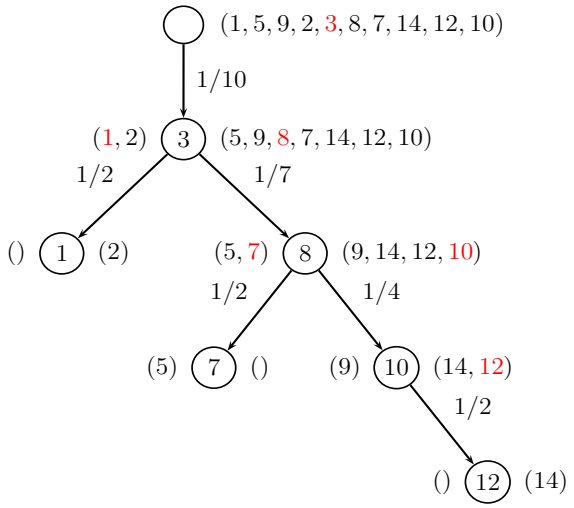

Fig. 6.3 A tree representation of a run of quicksort

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 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 element of the corresponding list, say Y, was chosen uniformly, namely $1/|Y|$. So, given an input list S of length n, the sample space Ω is the set of all computation trees T with root label S. We assign a probability to the trees T in Ω as follows: If $n = 0,1$, 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 $\mathsf{Pr}(T)$. We leave it as an exercise to prove that the sum of the probabilities of all the trees in Ω is equal to 1.

A random variable of great interest on (Ω,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. 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 (Ω,Pr), a *random variable* is any function $X\colon \Omega \to \mathbb{R}$. For any real number $a \in \mathbb{R}$, we define $\mathsf{Pr}(X = a)$ as the probability

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

and $\mathsf{Pr}(X \leq a)$ as the probability

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

The function $f\colon \mathbb{R} \to [0,1]$ given by

$$f(a) = \mathsf{Pr}(X = a), \quad a \in \mathbb{R}$$

is the *probability mass function* of X, and the function $F\colon \mathbb{R} \to [0,1]$ given by

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

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.

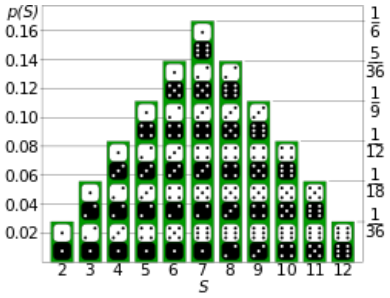

Fig. 6.4 The probability mass function for the sum of the numbers on two dice

If Ω is finite, then f only takes finitely many nonzero values; it is 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 $f(a) = \mathsf{Pr}(S = a)$.

The cumulative distribution function F has the following properties:

1. We have
$$\lim_{x \mapsto -\infty} F(x) = 0, \quad \lim_{x \mapsto \infty} F(x) = 1.$$
2. It is *monotonic nondecreasing*, which means that if $a \leq b$, then $F(a) \leq F(b)$.
3. It is piecewise constant with jumps, but it is *right-continuous*, which means that $\lim_{h>0, h \mapsto 0} F(a+h) = F(a)$.

For any $a \in \mathbb{R}$, because F is nondecreasing, we can define $F(a-)$ by

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

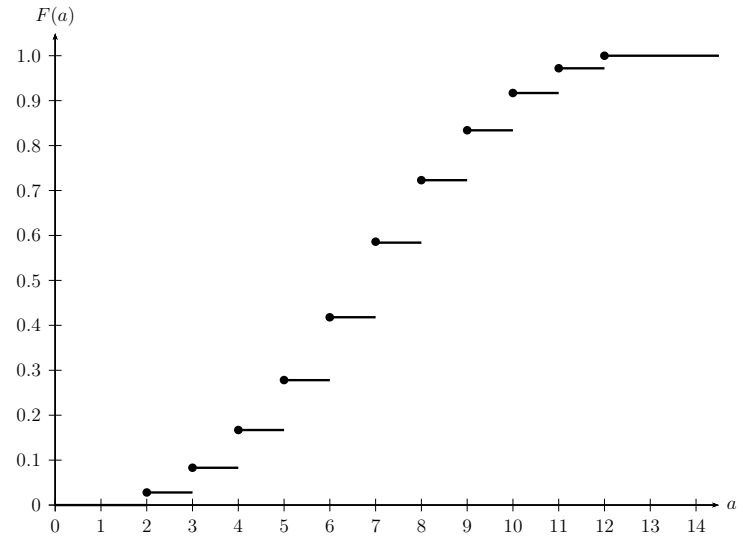

Fig. 6.5 The cumulative distribution function for the sum of the numbers on two dice

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

$$f(a) = \mathsf{Pr}(X = a) = F(a) - F(a-),$$

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

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

As a consequence, some of the definitions given in the discrete case in terms of the probabilities $\mathsf{Pr}(X = x)$, for example Definition 6.7, become trivial. These definitions need to be modifed; replacing $\mathsf{Pr}(X = x)$ by $\mathsf{Pr}(X \leq x)$ 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 Ω 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 Ω is the set of strings of length n over the alphabet $\{\mathrm{H}, \mathrm{T}\}$. 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 ω. Then, let X be the random variable defined such that $X(\omega)$ is the number of heads in ω. 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 ω 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 *binomial distribution*.

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 Ω is the infinite set of strings over the alphabet $\{\mathrm{H},\mathrm{T}\}$ of the form

$$\Omega = \{\mathrm{H},\mathrm{TH},\mathrm{TTH},\ldots,\ \mathrm{T}^n\mathrm{H},\ldots,\}.$$

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 ω. 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 *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 of failure (such trials that result exactly two outcomes, success or failure, are known as *Bernoulli trials*). Such processes are named after Jacob Bernoulli, a very significant contributor to probability theory after Fermat and Pascal.

Fig. 6.6 Jacob (Jacques) Bernoulli (1654–1705)

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 λ of "moderate" size. Let us show that we can approximate the distribution f of X in an interesting way. Indeed, for every nonnegative integer i, we can write

$$\begin{aligned} f(i) &= \binom{n}{i} p^i (1-p)^{n-i} \\ &= \frac{n!}{i!(n-i)!} \left(\frac{\lambda}{n}\right)^i \left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-i}. \end{aligned}$$

Now, for n large and λ moderate, we have

$$\left(1-\frac{\lambda}{n}\right)^{n} \approx \mathrm{e}^{-\lambda} \qquad \left(1-\frac{\lambda}{n}\right)^{-i} \approx 1 \qquad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1,$$

so we obtain

$$f(i) \approx \mathrm{e}^{-\lambda}\frac{\lambda^i}{i!}, \quad i \in \mathbb{N}.$$

The above is called a *Poisson distribution* with parameter λ. It is named after the French mathematician Simeon Denis Poisson.

Fig. 6.7 Siméon Denis Poisson (1781–1840)

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 $\mathsf{Pr}(A \mid B)$ 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 $\mathsf{Pr}(B) \neq 0$.

Definition 6.5. Given a discrete probability space (Ω, Pr), for any two events A and B, if $\mathsf{Pr}(B) \neq 0$, then we define the *conditional probability* $\mathsf{Pr}(A \mid B)$ 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

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{3}{36} \Big/ \frac{6}{36} = \frac{1}{2}.$$

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

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

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

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

of having at least one boy, and

$$A = \{\mathrm{BB}\}$$

of having two boys. We get

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

and so

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1}{4} \Big/ \frac{3}{4} = \frac{1}{3}.$$

Contrary to the popular belief that $\Pr(A \mid B) = 1/2$, it is actually equal to $1/3$. Now, consider the question: what is the probability that both are boys given that the first child is a boy? The answer to this question is indeed $1/2$.

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 doors?

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 $1/2$. 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 *Pi* for "the prize is behind door *i*," with $i = 1,2,3$, and Dj for "the host opens door Dj," for $j = 2,3$. Here, it is not necessary to consider the choice $D1$ since a sensible host will never open door 1. We can represent the sequences of choices occurrring in the game by a tree known as *probability tree* or *tree of possibilities*, shown in Figure 6.8.

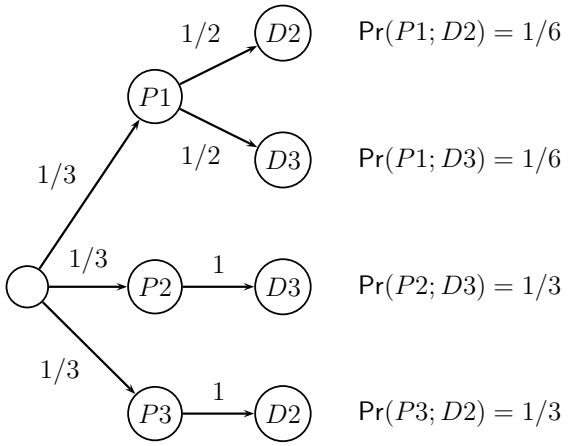

Fig. 6.8 The tree of possibilities in the Monty Hall problem

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

$$\Omega = \{P1;D2, P1;D3, P2;D3, P3;D2\}.$$

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

$$\Pr(P1;D2) = \frac{1}{6} \quad \Pr(P1;D3) = \frac{1}{6} \quad \Pr(P2;D3) = \frac{1}{3} \quad \Pr(P3;D2) = \frac{1}{3}.$$

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, $A = \{P1; D2, P1; D3\}$.
2. The prize is behind door 3; that is, $B = \{P3; D2\}$.
3. The host reveals door 2; that is, $C = \{P1; D2, P3; D2\}$.

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

1. $\mathsf{Pr}(A \mid C)$: the prize is behind door 1, given that the host reveals door 2.
2. $\mathsf{Pr}(B \mid C)$: the prize is behind door 3, given that the host reveals door 2.

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

$$\mathsf{Pr}(A \cap C) = 1/6,$$

and

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

so

$$\mathsf{Pr}(A \mid C) = \frac{\mathsf{Pr}(A \cap C)}{\mathsf{Pr}(C)} = \frac{1}{6} \Big/ \frac{1}{2} = \frac{1}{3}.$$

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

$$\mathsf{Pr}(B \cap C) = 1/3,$$

and

$$\mathsf{Pr}(B \mid C) = \frac{\mathsf{Pr}(B \cap C)}{\mathsf{Pr}(C)} = \frac{1}{3} \Big/ \frac{1}{2} = \frac{2}{3}.$$

Since $2/3 > 1/3$, the contestant has a greater chance (twice as big) to win the bigger prize by switching doors. 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 doors. 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 orginally 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 6.3. *(Bayes' Rules) For any two events A, B with $\mathsf{Pr}(A) > 0$ and $\mathsf{Pr}(B) > 0$, we have the following formulae:*

1. (Bayes' rule of retrodiction)

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

2. (Bayes' rule of exclusive and exhaustive clauses) If we also have $\mathsf{Pr}(A) < 1$ *and* $\mathsf{Pr}(B) < 1$*, then*

$$\mathsf{Pr}(A) = \mathsf{Pr}(A \mid B)\mathsf{Pr}(B) + \mathsf{Pr}(A \mid \overline{B})\mathsf{Pr}(\overline{B}).$$

More generally, if $B_1, \ldots, B_n$ *form a partition of* Ω *with* $\mathsf{Pr}(B_i) > 0$ *(*$n \geq 2$*), then*

$$\mathsf{Pr}(A) = \sum_{i=1}^{n} \mathsf{Pr}(A \mid B_i)\mathsf{Pr}(B_i).$$

3. (Bayes' sequential formula) For any sequence of events $A_1, \ldots, A_n$*, we have*

$$\mathsf{Pr}\left(\bigcap_{i=1}^{n} A_i\right) = \mathsf{Pr}(A_1)\mathsf{Pr}(A_2 \mid A_1)\mathsf{Pr}(A_3 \mid A_1 \cap A_2) \cdots \mathsf{Pr}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right).$$

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

$$\begin{aligned} \mathsf{Pr}(A) &= \mathsf{Pr}(A \cap B) \cup \mathsf{Pr}(A \cap \overline{B}) \\ &= \mathsf{Pr}(A \mid B)\mathsf{Pr}(A) \cup \mathsf{Pr}(A \mid \overline{B})\mathsf{Pr}(\overline{B}). \end{aligned}$$

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

$$\mathsf{Pr}(B \mid A) = \frac{\mathsf{Pr}(A \mid B)\mathsf{Pr}(B)}{\mathsf{Pr}(A \mid B)\mathsf{Pr}(B) + \mathsf{Pr}(A \mid \overline{B})\mathsf{Pr}(\overline{B})},$$

also known as *Bayes' law*.

Bayes' rule of retrodiction is at the heart of the so-called *Bayesian framewok*. In this framework, one thinks of B as an event describing some state (such as having a certain desease) and of A an an event describing some measurement or test (such as having high blood pressure). One wishes to infer the *a posteriori* probability $\mathsf{Pr}(B \mid A)$ of the state B given the test A, in terms of the *prior* probability $\mathsf{Pr}(B)$ and the *likelihood function* $\mathsf{Pr}(A \mid B)$. The likelihood function $\mathsf{Pr}(A \mid B)$ is a measure of the likelihood of the test A given that we know the state B, and $\mathsf{Pr}(B)$ is a measure of our prior knowledge about the state; for example, having a certain disease. The

probability $\mathsf{Pr}(A)$ is usually obtained using Bayes's second rule because we also know $\mathsf{Pr}(A \mid \overline{B})$.

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 desease, 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 desease. What is the probability for a patient with a positive test to be affected by the desease?

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

$$\begin{aligned}\mathsf{Pr}(S) &= 0.001\\ \mathsf{Pr}(+ \mid S) &= 0.99\\ \mathsf{Pr}(+ \mid \overline{S}) &= 0.02,\end{aligned}$$

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

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

We also have

$$\mathsf{Pr}(+) = \mathsf{Pr}(+ \mid S)\mathsf{Pr}(S) + \mathsf{Pr}(+ \mid \overline{S})\mathsf{Pr}(\overline{S}),$$

so we obtain

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

Since this probability is small, one is led to question the reliability of 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

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

Redo the calculations with the new data

$$\begin{aligned}\mathsf{Pr}(S) &= 0.00001\\ \mathsf{Pr}(+ \mid S) &= 0.99\\ \mathsf{Pr}(+ \mid \overline{S}) &= 0.01.\end{aligned}$$

You will find that the probability $\mathsf{Pr}(S \mid +)$ is approximately 0.000099, so the chance of being sick is rather small, and it is more likely that the test was incorrect.

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

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

Asuming that $\mathsf{Pr}(A) \neq 0$ and $\mathsf{Pr}(B) \neq 0$, we have

$$\mathsf{Pr}(A \cap B) = \mathsf{Pr}(A \mid B)\mathsf{Pr}(B) = \mathsf{Pr}(B \mid A)\mathsf{Pr}(A),$$

so we get the following proposition.

Proposition 6.4. *For any two events A, B such that $\mathsf{Pr}(A) \neq 0$ and $\mathsf{Pr}(B) \neq 0$, the following statements are equivalent:*

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

Remark: For a fixed event B with $\mathsf{Pr}(B) > 0$, the function $A \mapsto \mathsf{Pr}(A \mid B)$ satisfies the axioms of a probability measure stated in Definition 6.2. This is shown in Ross [11] (Section 3.5), among other references.

The examples where we flip a coin n times or roll two dice n times are examples of *indendent repeated trials*. They suggest the following definition.

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

$$\mathsf{Pr}(\omega_1, \omega_2) = \mathsf{Pr}_1(\omega_1)\mathsf{Pr}_2(\Omega_2), \quad \omega_1 \in \Omega_1, \omega_2 \in \Omega_2.$$

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

$$\mathsf{Pr}(\omega_1, \ldots, \omega_n) = \mathsf{Pr}(\omega_1) \cdots \mathsf{Pr}(\omega_n), \quad \omega_i \in \Omega, i = 1, \ldots, n.$$

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 $(\Omega_1, \mathscr{F}_1, \mathsf{Pr}_1)$ and $(\Omega_2, \mathscr{F}_2, \mathsf{Pr}_2)$ can also be defined, but $\mathscr{F}_1 \times \mathscr{F}_2$ is not a σ-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

$$f_{X,Y}(a,b) = \mathsf{Pr}(X = a \text{ and } Y = b) = \mathsf{Pr}(\{\omega \in \Omega \mid (X(\omega) = a) \wedge (Y(\omega) = b)\}),$$

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

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

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

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

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

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

the random variables S_1 and S_2 are independent.

Example 6.16. Suppose we flip a biased coin (with probability p of success) once. 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

$$\mathsf{Pr}(X = 1 \text{ and } Y = 1) = 0,$$

yet

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

Now, if we flip the coin N times, where N has the Poisson distribution with parameter λ, it is remarkable that X and Y are independent; see Grimmett and Stirzaker [6] (Section 3.2).

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

Proposition 6.5. *If X and Y are two random variables on a discrete probability space (Ω, 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*

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

Given the joint mass function $f_{X,Y}$ of two random variables X and Y, the mass functions f_X of X and f_Y of Y are called *marginal mass functions*, and they are obtained from $f_{X,Y}$ by the formulae

$$f_X(x) = \sum_y f_{X,Y}(x,y), \quad f_Y(y) = \sum_x f_{X,Y}(x,y).$$

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

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

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

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

They correspond to the *marginal distribution functions* F_X of X and F_Y of Y given by

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

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 in 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 $\mathsf{Pr}(X = a_j)$ 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

$$a_1\mathsf{Pr}(X=a_1)+a_2\mathsf{Pr}(X=a_2)+\cdots+a_n\mathsf{Pr}(X=a_n).$$

Definition 6.8. Given a discrete probability space (Ω, Pr), for any random variable X, the *mean value* or *expected value* or *expectation*[1] of X is the number $\mathsf{E}(X)$ defined as

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

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

$$\mathsf{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathsf{Pr}(\omega).$$

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

$$\mathsf{Pr}(X\le x)\ge\frac{1}{2} \quad \text{and} \quad \mathsf{Pr}(X\ge x)\ge\frac{1}{2}.$$

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

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

provided that the above sum converges absolutely (that is, the partial sums of absolute values converge). If we have a probability space $(X, \mathscr{F}, \mathsf{Pr})$ with Ω 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

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

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

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

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

[1] It is amusing that in French, the word for *expectation* is *espérance mathématique*. There is hope for mathematics!

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

$$\mathsf{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathsf{Pr}(\omega) = c \sum_{\omega \in \Omega} \mathsf{Pr}(\omega) = c\mathsf{Pr}(\Omega) = c,$$

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

Example 6.17. Consider the sum S of the values on the dice from Example 6.6. The expectation of S is

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

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

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

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 $\mathsf{E}(X)$. 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

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

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

$$\mathsf{E}(X) = np.$$

It should be observed that the expectation of a random variable may be infinite. For example, if X is a random variable whose probability mass function f is given by

$$f(k) = \frac{1}{k(k+1)}, \quad k = 1, 2, \ldots,$$

then $\sum_{k\in\mathbb{N}-\{0\}} f(k) = 1$, since

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \sum_{k=1}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1}\right) = \lim_{k\mapsto\infty} \left(1 - \frac{1}{k+1}\right) = 1,$$

but

$$\mathsf{E}(X) = \sum_{k\in\mathbb{N}-\{0\}} kf(k) = \sum_{k\in\mathbb{N}-\{0\}} \frac{1}{k+1} = \infty.$$

A crucial property of expectation that often allows simplifications in computing the expectation of a random variable is its linearity.

Proposition 6.6. *(Linearity of Expectation) Given two random variables on a discrete probability space, for any real number λ, we have*

$$\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$$
$$\mathsf{E}(\lambda X) = \lambda \mathsf{E}(X).$$

Proof. We have

$$\begin{aligned}
\mathsf{E}(X+Y) &= \sum_z z \cdot \mathsf{Pr}(X+Y=z) \\
&= \sum_x \sum_y (x+y) \cdot \mathsf{Pr}(X=x \text{ and } Y=y) \\
&= \sum_x \sum_y x \cdot \mathsf{Pr}(X=x \text{ and } Y=y) + \sum_x \sum_y y \cdot \mathsf{Pr}(X=x \text{ and } Y=y) \\
&= \sum_x \sum_y x \cdot \mathsf{Pr}(X=x \text{ and } Y=y) + \sum_y \sum_x y \cdot \mathsf{Pr}(X=x \text{ and } Y=y) \\
&= \sum_x x \sum_y \mathsf{Pr}(X=x \text{ and } Y=y) + \sum_y y \sum_x \mathsf{Pr}(X=x \text{ and } Y=y).
\end{aligned}$$

Now, the events $A_x = \{x \mid X = x\}$ form a partition of Ω, which implies that

$$\sum_y \mathsf{Pr}(X=x \text{ and } Y=y) = \mathsf{Pr}(X=x).$$

Similarly the events $B_y = \{y \mid Y = y\}$ form a partition of Ω, which implies that

$$\sum_x \mathsf{Pr}(X = x \text{ and } Y = y) = \mathsf{Pr}(Y = y).$$

By substitution, we obtain

$$\mathsf{E}(X+Y) = \sum_x x \cdot \mathsf{Pr}(X = x) + \sum_y y \cdot \mathsf{Pr}(Y = y),$$

proving that $\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$. When Ω 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

$$\begin{aligned}
\mathsf{E}(\lambda X) &= \sum_x x \cdot \mathsf{Pr}(\lambda X = x) \\
&= \lambda \sum_x \frac{x}{\lambda} \cdot \mathsf{Pr}(X = x/\lambda) \\
&= \lambda \sum_x y \cdot \mathsf{Pr}(X = y) \\
&= \lambda \mathsf{E}(X).
\end{aligned}$$

as claimed. If $\lambda = 0$, the equation is trivial. □

By a trivial induction, we obtain that for any finite number of random variables $X_1, \ldots, X_n$, we have

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

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 ith 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 $\mathsf{E}(X_i) = p$, and since

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

we get

$$\mathsf{E}(X) = np.$$

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 Ω, 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 probabilty $\mathsf{Pr}(A)$ of the event A. Indeed,

$$\begin{aligned} \mathsf{E}(I_A) &= \sum_{\omega \in \Omega} I_A(\omega)\mathsf{Pr}(\omega) \\ &= \sum_{\omega \in A} \mathsf{Pr}(\omega) \\ &= \mathsf{Pr}(A). \end{aligned}$$

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 Ω be the set of all $n!$ permutations of $\{1, \ldots, n\}$, and give Ω the uniform probabilty measure; that is, for every permutation π, let

$$\mathsf{Pr}(\pi) = \frac{1}{n!}.$$

We say that these are *random permutations*. A *fixed point* of a permutation π 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 π. 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

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

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

$$\mathsf{E}(X_k) = \mathsf{Pr}(X_k = 1).$$

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

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

On average, a random permutation has one fixed point.

If X is a random variable on a discrete probability space Ω (possibly countably infinite), for any function $g\colon \mathbb{R} \to \mathbb{R}$, 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 φ and ψ are two functions, we leave it as an exercise to prove that if X and Y are independent, then so are $\varphi(X)$ and $\psi(Y)$.

Altough 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 6.7. *If X is a random variable on a discrete probability space Ω, for any function $g\colon \mathbb{R} \to \mathbb{R}$, the expectation $\mathsf{E}(g(X))$ of $g(X)$ (if it exists) is given by*

$$\mathsf{E}(g(X)) = \sum_x g(x) f(x),$$

where f is the mass function of X.

Proof. We have

$$\begin{aligned}
\mathsf{E}(g(X)) &= \sum_y y \cdot \mathsf{Pr}(g \circ X = y) \\
&= \sum_y y \cdot \mathsf{Pr}(\{\omega \in \Omega \mid g(X(\omega)) = y\}) \\
&= \sum_y y \sum_x \mathsf{Pr}(\{\omega \in \Omega \mid g(x) = y, X(\omega) = x\}) \\
&= \sum_y \sum_{x, g(x)=y} y \cdot \mathsf{Pr}(\{\omega \in \Omega, \mid X(\omega) = x\}) \\
&= \sum_y \sum_{x, g(x)=y} g(x) \cdot \mathsf{Pr}(X = x) \\
&= \sum_x g(x) \cdot \mathsf{Pr}(X = x) \\
&= \sum_x g(x) f(x),
\end{aligned}$$

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.

Given two random variables X and Y on a discrete probability space Ω, for any function $g\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, then $g(X,Y)$ is a random variable and it is easy to show that $\mathsf{E}(g(X,Y))$ (if it exists) is given by

$$\mathsf{E}(g(X,Y)) = \sum_{x,y} g(x,y) f_{X,Y}(x,y),$$

where $f_{X,Y}$ is the joint mass function of X and Y.

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 $\mathsf{E}(X^k)$, for $k \geq 2$. We have

$$\begin{aligned}\mathsf{E}(X^k) &= \sum_{i=0}^{n} i^k f(i) \\ &= \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}.\end{aligned}$$

Recall that

$$i\binom{n}{i} = n\binom{n-1}{i-1}.$$

Using this, we get

$$\begin{aligned}\mathsf{E}(X^k) &= \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\ &= np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \quad (\text{let } j = i-1) \\ &= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\ &= np\mathsf{E}((Y+1)^{k-1}),\end{aligned}$$

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 $\mathsf{E}(X^k)$. For $k = 2$, we get

$$\mathsf{E}(X^2) = np\mathsf{E}(Y+1) = np((n-1)p+1).$$

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

Proposition 6.8. *If X is a random variable that takes on only nonnegative integers, then its expectation $\mathsf{E}(X)$ (if it exists) is given by*

$$\mathsf{E}(X) = \sum_{i=1}^{\infty} \mathsf{Pr}(X \geq i).$$

Proof. For any integer $n \geq 1$, we have

$$\sum_{j=1}^{n} j\mathsf{Pr}(X=j) = \sum_{j=1}^{n} \sum_{i=1}^{j} \mathsf{Pr}(X=j) = \sum_{i=1}^{n} \sum_{j=i}^{n} \mathsf{Pr}(X=j) = \sum_{i=1}^{n} \mathsf{Pr}(n \geq X \geq i).$$

Then, if we let n go to infinity, we get

$$\sum_{i=1}^{\infty} \mathsf{Pr}(X \geq i) = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \mathsf{Pr}(X = j) = \sum_{j=1}^{\infty} \sum_{i=1}^{j} \mathsf{Pr}(X = j) = \sum_{j=1}^{\infty} j\mathsf{Pr}(X = j) = \mathsf{E}(X),$$

as claimed. □

Proposition 6.8 has the following intuitive geometric interpretation: $\mathsf{E}(X)$ is the area above the graph of the cumulative distribution function $F(i) = \mathsf{Pr}(X \leq i)$ of X and below the horizontal line $F = 1$. Here is an application of Proposition 6.8.

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 6.8. We have

$$\begin{aligned}
\mathsf{Pr}(X \geq i) &= \sum_{i=i}^{\infty} (1-p)^{i-1}p \\
&= p(1-p)^{i-1} \sum_{j=0}^{\infty} (1-p)^{j} \\
&= p(1-p)^{i-1} \frac{1}{1-(1-p)} \\
&= (1-p)^{i-1}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathsf{E}(X) &= \sum_{i=1}^{\infty} \mathsf{Pr}(X \geq i) \\
&= \sum_{i=1}^{\infty} (1-p)^{i-1}. \\
&= \frac{1}{1-(1-p)} = \frac{1}{p}.
\end{aligned}$$

Therefore,

$$\mathsf{E}(X) = \frac{1}{p},$$

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

Let us now compute $\mathsf{E}(X^2)$. We have

$$\begin{aligned}
\mathsf{E}(X^2) &= \sum_{i=1}^{\infty} i^2(1-p)^{i-1}p \\
&= \sum_{i=1}^{\infty} (i-1+1)^2(1-p)^{i-1}p \\
&= \sum_{i=1}^{\infty} (i-1)^2(1-p)^{i-1}p + \sum_{i=1}^{\infty} 2(i-1)(1-p)^{i-1}p + \sum_{i=1}^{\infty} (1-p)^{i-1}p \\
&= \sum_{j=0}^{\infty} j^2(1-p)^j p + 2\sum_{j=1}^{\infty} j(1-p)^j p + 1 \quad (\text{let } j = i-1) \\
&= (1-p)\mathsf{E}(X^2) + 2(1-p)\mathsf{E}(X) + 1.
\end{aligned}$$

Since $\mathsf{E}(X) = 1/p$, we obtain

$$\begin{aligned}
p\mathsf{E}(X^2) &= \frac{2(1-p)}{p} + 1 \\
&= \frac{2-p}{p},
\end{aligned}$$

so

$$\mathsf{E}(X^2) = \frac{2-p}{p^2}.$$

By the way, the trick of writing $i = i-1+1$ can be used to compute $\mathsf{E}(X)$. Try to recompute $\mathsf{E}(X)$ 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 $\mathsf{E}(X)$, we decompose X as a sum of indicator variables $X_{i,j}$, with $X_{i,j} = 1$ iff y_i and y_j are ever compared, and $X_{i,j} = 0$ otherwise. Then, it is clear that

$$X = \sum_{j=2}^{n} \sum_{i=1}^{j-1} X_{i,j},$$

and

$$\mathsf{E}(X) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathsf{E}(X_{i,j}).$$

Furthermore, since $X_{i,j}$ is an indicator variable, we have

$$\mathsf{E}(X_{i,j}) = \mathsf{Pr}(y_i \text{ and } y_j \text{ are ever compared}).$$

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

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

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 mutally exclusive, we have

$$\mathsf{E}(X_{i,j}) = \mathsf{Pr}(y_i \text{ and } y_j \text{ are ever compared}) = \frac{2}{j-i+1}.$$

It follows that

$$\begin{aligned}
\mathsf{E}(X) &= \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathsf{E}(X_{i,j}) \\
&= 2 \sum_{j=2}^{n} \sum_{k=2}^{j} \frac{1}{k} \quad (\text{set } k = j-i+1) \\
&= 2 \sum_{k=2}^{n} \sum_{j=k}^{n} \frac{1}{k} \\
&= 2 \sum_{k=2}^{n} \frac{n-k+1}{k} \\
&= 2(n+1) \sum_{k=1}^{n} \frac{1}{k} - 4n.
\end{aligned}$$

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

$$\ln(n) + \frac{1}{n} \leq H_n \leq \ln n + 1.$$

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

$$\mathsf{E}(X) = 2n\ln + \Theta(n).$$

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

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

$$\mathsf{E}(X) = \lambda.$$

Recall that a Poisson distribution is given by

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

so we have

$$\begin{aligned}\mathsf{E}(X) &= \sum_{i=0}^{\infty} i\mathrm{e}^{-\lambda}\frac{\lambda^i}{i!}\\ &= \lambda\mathrm{e}^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}\\ &= \lambda\mathrm{e}^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} \quad (\text{let } j = i-1)\\ &= \lambda\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = \lambda,\end{aligned}$$

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

$$\mathsf{E}(X^2) = \lambda(\lambda+1).$$

Alhough in general $\mathsf{E}(XY) \neq \mathsf{E}(X)\mathsf{E}(Y)$, this is true for independent random variables.

Proposition 6.9. *If two random variables X and Y on the same discrete probability space are independent, then*

$$\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y).$$

Proof. We have

$$\begin{aligned}
\mathsf{E}(XY) &= \sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathsf{Pr}(\omega) \\
&= \sum_x \sum_y xy \cdot \mathsf{Pr}(X = x \text{ and } Y = y) \\
&= \sum_x \sum_y xy \cdot \mathsf{Pr}(X = x)\mathsf{Pr}(Y = y) \\
&= \left(\sum_x x \cdot \mathsf{Pr}(X = x)\right)\left(\sum_y y \cdot \mathsf{Pr}(Y = y)\right) \\
&= \mathsf{E}(X)\mathsf{E}(Y),
\end{aligned}$$

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

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

$$\mathsf{E}(P) = \mathsf{E}(S_1)\mathsf{E}(S_2) = \frac{7}{2} \cdot \frac{7}{2} = \frac{49}{4},$$

since $\mathsf{E}(S_1) = \mathsf{E}(S_2) = 7/2$, as we easily determine since all probabilities are equal to $1/6$. On the other hand, S and P are not independent (check it).

6.5 Variance, Standard Deviation, Chebyshev's Inequality

The mean (expectation) $\mathsf{E}(X)$ of a random variable X gives some useful information about it, but it does not say how X is spread. Another quantity, the *variance* $\mathsf{Var}(X)$, measure the spread of the distribution by finding the "average" of the square difference $(X - \mathsf{E}(X))^2$, namely

$$\mathsf{Var}(X) = \mathsf{E}(X - \mathsf{E}(X))^2.$$

Note that computing $\mathsf{E}(X - \mathsf{E}(X))$ yields no information since

$$\mathsf{E}(X - \mathsf{E}(X)) = \mathsf{E}(X) - \mathsf{E}(\mathsf{E}(X)) = \mathsf{E}(X) - \mathsf{E}(X) = 0.$$

Definition 6.10. Given a discrete probability space (Ω, Pr), for any random variable X, the *variance* $\mathsf{Var}(X)$ of X (if it exists) is defined as

$$\mathsf{Var}(X) = \mathsf{E}(X - \mathsf{E}(X))^2.$$

The expectation $\mathsf{E}(X)$ of a random variable X is often denoted by μ. The variance is also denoted $\mathsf{V}(X)$, for instance, in Graham, Knuth and Patashnik [5]).

Since the variance $\mathsf{Var}(X)$ involves a square, it can be quite large, so it is convenient to take its square root and to define the *standard deviation* of σ X as

$$\sigma = \sqrt{\mathsf{Var}(X)}.$$

The following result shows that the variance $\mathsf{Var}(X)$ can be computed using $\mathsf{E}(X^2)$ and $\mathsf{E}(X)$.

Proposition 6.10. *Given a discrete probability space* (Ω, Pr)*, for any random variable* X*, the variance* $\mathsf{Var}(X)$ *of* X *is given by*

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2.$$

Consequently, $\mathsf{Var}(X) \leq \mathsf{E}(X^2)$.

Proof. Using the linearity of expectation and the fact that the expectation of a constant is itself, we have

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{E}(X - \mathsf{E}(X))^2 \\
&= \mathsf{E}(X^2 - 2X\mathsf{E}(X) + (\mathsf{E}(X))^2) \\
&= \mathsf{E}(X^2) - 2\mathsf{E}(X)\mathsf{E}(X) + (\mathsf{E}(X))^2 \\
&= \mathsf{E}(X^2) - (\mathsf{E}(X))^2
\end{aligned}$$

as claimed. □

For example, if we roll a fair dice, we know that the number S_1 on the dice has expectation $\mathsf{E}(S_1) = 7/2$. We also have

$$\mathsf{E}(S_1^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6},$$

so the variance of S_1 is

$$\mathsf{Var}(S_1) = \mathsf{E}(S_1^2) - (\mathsf{E}(S_1))^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

The quantity $\mathsf{E}(X^2)$ is called the *second moment* of X. More generally, we have the following definition.

Definition 6.11. Given a random variable X on a discrete probability space (Ω, Pr), for any integer $k \geq 1$, the *kth moment* μ_k of X is given by $\mu_k = \mathsf{E}(X^k)$, and the *kth central moment* σ_k of X is defined by $\sigma_k = \mathsf{E}((X - \mu_1)^k)$.

Typically, only $\mu = \mu_1$ and σ_2 are of interest. As before, $\sigma = \sqrt{\sigma_2}$. However, σ_3 and σ_4 give rise to quantities with exotic names: the *skewness* (σ_3/σ^3) and the *kurtosis* $(\sigma_4/\sigma^4 - 3)$.

We can easily compute the variance of a random variable for the binomial distribution and the geometric distribution, since we already computed $\mathsf{E}(X^2)$.

Example 6.25. In Example 6.21, the case of a binomial distribution, we found that

$$\mathsf{E}(X^2) = np\mathsf{E}(Y+1) = np((n-1)p+1).$$

We also found earlier (Example 6.19) that $\mathsf{E}(X) = np$. Therefore, we have

$$\begin{aligned}\mathsf{Var}(X) &= \mathsf{E}(X^2) - (\mathsf{E}(X))^2 \\ &= np((n-1)p+1) - (np)^2 \\ &= np(1-p).\end{aligned}$$

Therefore,

$$\mathsf{Var}(X) = np(1-p).$$

Example 6.26. In Example 6.22, the case of a geometric distribution, we found that

$$\begin{aligned}\mathsf{E}(X) &= \frac{1}{p} \\ \mathsf{E}(X^2) &= \frac{2-p}{p^2}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathsf{Var}(X) &= \mathsf{E}(X^2) - (\mathsf{E}(X))^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}.\end{aligned}$$

Therefore,

$$\mathsf{Var}(X) = \frac{1-p}{p^2}.$$

Example 6.27. In Example 6.24, the case of a Poisson distribution with parameter λ, we found that

$$\begin{aligned}\mathsf{E}(X) &= \lambda \\ \mathsf{E}(X^2) &= \lambda(\lambda+1).\end{aligned}$$

It follows that

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

Therefore, a random variable with a Poisson distribution has the same value for its expectation and its variance,

$$\mathsf{E}(X) = \mathsf{Var}(X) = \lambda.$$

In general, if X and Y are not independent variables, $\mathsf{Var}(X+Y) \neq \mathsf{Var}(X) + \mathsf{Var}(Y)$. However, if they are, things are great!

Proposition 6.11. *Given a discrete probability space (Ω, Pr), for any random variable X and Y, if X and Y are independent, then*

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y).$$

Proof. Recall from Proposition 6.9 that if X and Y are independent, then $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$. Then, we have

$$\begin{aligned}\mathsf{E}((X+Y)^2) &= \mathsf{E}(X^2 + 2XY + Y^2)\\ &= \mathsf{E}(X^2) + 2\mathsf{E}(XY) + \mathsf{E}(Y^2)\\ &= \mathsf{E}(X^2) + 2\mathsf{E}(X)\mathsf{E}(Y) + \mathsf{E}(Y^2).\end{aligned}$$

Using this, we get

$$\begin{aligned}\mathsf{Var}(X+Y) &= \mathsf{E}((X+Y)^2) - (\mathsf{E}(X+Y))^2\\ &= \mathsf{E}(X^2) + 2\mathsf{E}(X)\mathsf{E}(Y) + \mathsf{E}(Y^2) - ((\mathsf{E}(X))^2 + 2\mathsf{E}(X)\mathsf{E}(Y) + (\mathsf{E}(Y))^2)\\ &= \mathsf{E}(X^2) - (\mathsf{E}(X))^2 + \mathsf{E}(Y^2) - (\mathsf{E}(Y))^2\\ &= \mathsf{Var}(X) + \mathsf{Var}(Y),\end{aligned}$$

as claimed. □

The following proposition is also useful.

Proposition 6.12. *Given a discrete probability space (Ω, Pr), for any random variable X, the following properties hold:*

1. *If $X \geq 0$, then $\mathsf{E}(X) \geq 0$.*
2. *If X is a random variable with constant value λ, then $\mathsf{E}(X) = \lambda$.*
3. *For any two random variables X and Y defined on the probablity space (Ω, Pr), if $X \leq Y$, which means that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\mathsf{E}(X) \leq \mathsf{E}(Y)$ (monotonicity of expectation).*
4. *For any scalar $\lambda \in \mathbb{R}$, we have*

$$\mathsf{Var}(\lambda X) = \lambda^2 \mathsf{Var}(X).$$

Proof. Properties (1) and (2) are obvious. For (3), $X \leq Y$ iff $Y - X \geq 0$, so by (1) we have $\mathsf{E}(Y-X) \geq 0$, and by linearity of expectation, $\mathsf{E}(Y) \geq \mathsf{E}(X)$. For (4), we have

$$\begin{aligned}\mathsf{Var}(\lambda X) &= \mathsf{E}\big((\lambda X - \mathsf{E}(\lambda X))^2\big)\\ &= \mathsf{E}\big(\lambda^2 (X - \mathsf{E}(X))^2\big)\\ &= \lambda^2 \mathsf{E}\big((X - \mathsf{E}(X))^2\big) = \lambda^2 \mathsf{Var}(X),\end{aligned}$$

as claimed. □

Property (4) shows that unlike expectation, the variance is not linear (although for independent random variables, $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$. This also holds in the more general case of uncorrelated random variables; see Proposition 6.13 below).

As an application of Proposition 6.11, if we consider the event of rolling two dice, since we showed that the random variables S_1 and S_2 are independent, we can compute the variance of their sum $S = S_1 + S_2$ and we get

$$\mathsf{Var}(S) = \mathsf{Var}(S_1) + \mathsf{Var}(S_2) = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}.$$

Recall that $\mathsf{E}(S) = 7$.

Here is an application of geometrically distributed random variables.

Example 6.28. Suppose there are m different types of coupons (or perhaps, the kinds of cards that kids like to collect), and that each time one obtains a coupon, it is equally likely to be any of these types. Let X denote the number of coupons one needs to collect in order to have at least one of each type. What is the expected value $\mathsf{E}(X)$ of X? This problem is usually called a *coupon collecting problem*.

The trick is to introduce the random variables X_i, where X_i is the number of additional coupons needed, after i distinct types have been collected, until another new type is obtained, for $i = 0, 1, \ldots, m-1$. Clearly,

$$X = \sum_{i=0}^{m-1} X_i,$$

and each X_i has a geometric distribution, where each trial has probability of success $p_i = (m-i)/m$. We know (see Example 6.22,) that

$$\mathsf{E}(X_i) = \frac{1}{p_i} = \frac{m}{m-i}.$$

Consequently,

$$\mathsf{E}(X) = \sum_{i=0}^{m-1} \mathsf{E}(X_i) = \sum_{i=0}^{m-1} \frac{m}{m-i} = m \sum_{i=1}^{m} \frac{1}{i}.$$

Once again, the *harmonic number*

$$H_m = \sum_{k=1}^{m} \frac{1}{k}$$

shows up! Since $H_n = \ln n + \Theta(1)$, we obtain

$$\mathsf{E}(X) = m \ln m + \Theta(m).$$

For example, if $m = 50$, then $\ln 50 = 3.912$, and $m \ln m \approx 196$. If $m = 100$, then $\ln 100 = 4.6052$, and $m \ln m \approx 461$. If the coupons are expensive, one begins to see why the company makes money!

It turns out that using a little bit of analysis, we can compute the variance of X. This is because it is easy to check that the X_i are independent, so

$$\mathsf{Var}(X) = \sum_{i=0}^{m-1} \mathsf{Var}(X_i).$$

From Example 6.22, we have

$$\mathsf{Var}(X_i) = \frac{1-p_i}{p_i^2} = \left(1 - \frac{m-i}{m}\right) \Big/ \frac{m^2}{(m-i)^2} = \frac{mi}{(m-i)^2}.$$

It follows that

$$\mathsf{Var}(X) = \sum_{i=0}^{m-1} \mathsf{Var}(X_i) = m \sum_{i=1}^{m} \frac{i}{(m-i)^2}.$$

To compute this sum, write

$$\begin{aligned}
\sum_{i=0}^{m-1} \frac{i}{(m-i)^2} &= \sum_{i=0}^{m-1} \frac{m}{(m-i)^2} - \sum_{i=0}^{m-1} \frac{m-i}{(m-i)^2} \\
&= \sum_{i=0}^{m-1} \frac{m}{(m-i)^2} - \sum_{i=0}^{m-1} \frac{1}{(m-i)} \\
&= m \sum_{j=1}^{m} \frac{1}{j^2} - \sum_{j=1}^{m} \frac{1}{j}.
\end{aligned}$$

Now, it is well known from analysis that

$$\lim_{m \mapsto \infty} \sum_{j=1}^{m} \frac{1}{j^2} = \frac{\pi^2}{6},$$

so we get

$$\mathsf{Var}(X) = \frac{m^2 \pi^2}{6} + \Theta(m \ln m).$$

Let us go back to the example about fixed points of random permutations (Example 6.20). We found that the expectation of the number of fixed points is $\mu = 1$. The reader should compute the standard deviation. The difficulty is that the random variables X_k are not independent, (for every permutation π, we have $X_k(\pi) = 1$ iff $\pi(k) = k$, and 0 otherwise). You will find that $\sigma = 1$. If you get stuck, look at Graham, Knuth and Patashnik [5], Chapter 8.

If X and Y are not independent, we still have

$$\begin{aligned}\mathsf{E}((X+Y)^2) &= \mathsf{E}(X^2+2XY+Y^2)\\ &= \mathsf{E}(X^2)+2\mathsf{E}(XY)+\mathsf{E}(Y^2),\end{aligned}$$

and we get

$$\begin{aligned}\mathsf{Var}(X+Y) &= \mathsf{E}((X+Y)^2)-(\mathsf{E}(X+Y))^2\\ &= \mathsf{E}(X^2)+2\mathsf{E}(XY)+\mathsf{E}(Y^2)-((\mathsf{E}(X))^2+2\mathsf{E}(X)\mathsf{E}(Y)+(\mathsf{E}(Y))^2)\\ &= \mathsf{E}(X^2)-(\mathsf{E}(X))^2+\mathsf{E}(Y^2)-(\mathsf{E}(Y))^2+2(\mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y))\\ &= \mathsf{Var}(X)+\mathsf{Var}(Y)+2(\mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y)).\end{aligned}$$

The term $\mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y)$ has a more convenient form. Indeed, we have

$$\begin{aligned}\mathsf{E}\big((X-\mathsf{E}(X))(Y-\mathsf{E}(Y))\big) &= \mathsf{E}\big(XY-X\mathsf{E}(Y)-\mathsf{E}(X)Y+\mathsf{E}(X)\mathsf{E}(Y)\big)\\ &= \mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y)-\mathsf{E}(X)\mathsf{E}(Y)+\mathsf{E}(X)\mathsf{E}(Y)\\ &= \mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y).\end{aligned}$$

In summary we proved that

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X)+\mathsf{Var}(Y)+2\mathsf{E}\big((X-\mathsf{E}(X))(Y-\mathsf{E}(Y))\big).$$

The quantity $\mathsf{E}\big((X-\mathsf{E}(X))(Y-\mathsf{E}(Y))\big)$ is well known in probability theory.

Definition 6.12. Given two random variables X and Y, their *covariance* $\mathsf{Cov}(X,Y)$ is defined by

$$\mathsf{Cov}(X,Y) = \mathsf{E}\big((X-\mathsf{E}(X))(Y-\mathsf{E}(Y))\big) = \mathsf{E}(XY)-\mathsf{E}(X)\mathsf{E}(Y).$$

If $\mathsf{Cov}(X,Y)=0$ (equivalently if $\mathsf{E}(XY)=\mathsf{E}(X)\mathsf{E}(Y)$) we say that X and Y are *uncorrelated*.

Observe that the variance of X is expressed in terms of the covariance of X by

$$\mathsf{Var}(X) = \mathsf{Cov}(X,X).$$

Let us recap the result of our computation of $\mathsf{Var}(X+Y)$ in terms of $\mathsf{Cov}(X,Y)$ as the following proposition.

Proposition 6.13. *Given two random variables X and Y, we have*

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X)+\mathsf{Var}(Y)+2\mathsf{Cov}(X,Y).$$

Therefore, if X an Y are uncorrelated ($\mathsf{Cov}(X,Y)=0$*), then*

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X)+\mathsf{Var}(Y).$$

In particular, if X and Y are independent, then X and Y are uncorrelated because

$$\mathsf{Cov}(X,Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y) = \mathsf{E}(X)\mathsf{E}(Y) - \mathsf{E}(X)\mathsf{E}(Y) = 0.$$

This yields another proof of Proposition 6.11.

However, beware that $\mathsf{Cov}(X,Y) = 0$ does not necessarily imply that X and Y are independent. For example, let X and Y be the random variables defined on $\{-1,0,1\}$ by

$$\mathsf{Pr}(X=0) = \mathsf{Pr}(X=1) = \mathsf{Pr}(X=-1) = \frac{1}{3},$$

and

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0\,. \end{cases}$$

Since $XY = 0$, we have $\mathsf{E}(XY) = 0$, and since we also have $\mathsf{E}(X) = 0$, we have

$$\mathsf{Cov}(X,Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y) = 0.$$

However, the reader will check easily that X and Y are not independent.

A better measure of independence is given by the *correlation coefficient* $\rho(X,Y)$ of X and Y, given by

$$\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}},$$

provided that $\mathsf{Var}(X) \neq 0$ and $\mathsf{Var}(Y) \neq 0$. It turns out that $|\rho(X,Y)| \leq 1$, which is shown using the Cauchy–Schwarz inequality.

Proposition 6.14. *(Cauchy–Schwarz inequality) For any two random variables X and Y on a discrete probability space Ω, we have*

$$|\mathsf{E}(XY)| \leq \sqrt{\mathsf{E}(X^2)}\sqrt{\mathsf{E}(Y^2)}.$$

Equality is achieved if and only if there exist some $\alpha, \beta \in \mathbb{R}$ (not both zero) such that $\mathsf{E}((\alpha X + \beta Y)^2) = 0$.

Proof. This is a standard argument involving a quadratic equation. For any $\lambda \in \mathbb{R}$, define the function $T(\lambda)$ by

$$T(\lambda) = \mathsf{E}((X + \lambda Y)^2).$$

We get

$$\begin{aligned} T(\lambda) &= \mathsf{E}(X^2 + 2\lambda XY + \lambda^2 Y^2) \\ &= \mathsf{E}(X^2) + 2\lambda\,\mathsf{E}(XY) + \lambda^2\mathsf{E}(Y^2). \end{aligned}$$

Since $\mathsf{E}((X+\lambda Y)^2) \geq 0$, we have $T(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$. If $\mathsf{E}(Y^2) = 0$, then we must have $\mathsf{E}(XY) = 0$, since otherwise we could choose λ so that $\mathsf{E}(X^2) + 2\lambda\,\mathsf{E}(XY) < 0$. In this case, the inequality is trivial. If $\mathsf{E}(Y^2) > 0$, then for $T(\lambda)$ to be nonnegative the quadratic equation

$$\mathsf{E}(X^2) + 2\lambda\,\mathsf{E}(XY) + \lambda^2\mathsf{E}(Y^2) = 0$$

should have at most one real root, which is equivalent to the well-known condition

$$4(\mathsf{E}(XY))^2 - 4\mathsf{E}(X^2)\mathsf{E}(Y^2) \leq 0,$$

which is equivalent to

$$|\mathsf{E}(XY)| \leq \sqrt{\mathsf{E}(X^2)}\sqrt{\mathsf{E}(Y^2)},$$

as claimed. If $(\mathsf{E}(XY))^2 = \mathsf{E}(X^2)\mathsf{E}(Y^2)$, then either $\mathsf{E}(Y^2) = 0$, and then with $\alpha = 0, \beta = 1$, we have $\mathsf{E}((\alpha X + \beta Y)^2) = 0$, or $\mathsf{E}(Y^2) > 0$, in which case the quadratic equation

$$\mathsf{E}(X^2) + 2\lambda\,\mathsf{E}(XY) + \lambda^2\mathsf{E}(Y^2) = 0$$

has a unique real root λ_0, so we have $\mathsf{E}((X + \lambda_0 Y)^2) = 0$.

Conversely, if $\mathsf{E}((\alpha X + \beta Y)^2) = 0$ for some $\alpha, \beta \in \mathbb{R}$, then either $\mathsf{E}(Y^2) = 0$, in which case we showed that we also have $\mathsf{E}(XY) = 0$, or the quadratic equation has some real root, so we must have $(\mathsf{E}(XY))^2 - \mathsf{E}(X^2)\mathsf{E}(Y^2) = 0$. In both cases, we have $(\mathsf{E}(XY))^2 = \mathsf{E}(X^2)\mathsf{E}(Y^2)$. □

It can be shown that for any random variable Z, if $\mathsf{E}(Z^2) = 0$, then $\mathsf{Pr}(Z = 0) = 1$; see Grimmett and Stirzaker [6] (Chapter 3, Problem 3.11.2). In fact, this is a consequence of Proposition 6.2 and Chebyshev's Inequality (see below), as shown in Ross [11] (Section 8.2, Proposition 2.3). It follows that if equality is achieved in the Cauchy–Schwarz inequality, then there are some reals α, β (not both zero) such that $\mathsf{Pr}(\alpha X + \beta Y = 0) = 1$; in other words, X and Y are dependent with probability 1. If we apply the Cauchy-Schwarz inequality to the random variables $X - \mathsf{E}(X)$ and $Y - \mathsf{E}(Y)$, we obtain the following result.

Proposition 6.15. *For any two random variables X and Y on a discrete probability space, we have*

$$|\rho(X,Y)| \leq 1,$$

with equality iff there are some real numbers α, β, γ (with α, β not both zero) such that $\mathsf{Pr}(\alpha X + \beta Y = \gamma) = 1$.

As emphasized by Graham, Knuth and Patashnik [5], the variance plays a key role in an inquality due to Chebyshev (published in 1867) that tells us that a random variable will rarely be far from its mean $\mathsf{E}(X)$ if its variance $\mathsf{Var}(X)$ is small.

Proposition 6.16. *(Chebyshev's Inequality) If X is any random variable, for every $\alpha > 0$, we have*

$$\mathsf{Pr}\big((X - \mathsf{E}(X))^2 \geq \alpha\big) \leq \frac{\mathsf{Var}(X)}{\alpha}$$

Proof. We follow Knuth. We have

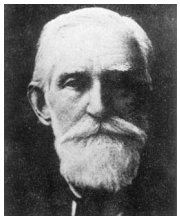

Fig. 6.9 Pafnuty Lvovich Chebyshev (1821–1894)

$$\begin{aligned}
\mathsf{Var}(X) &= \sum_{\omega\in\Omega} (X(\omega)-\mathsf{E}(X))^2\mathsf{Pr}(\omega) \\
&\geq \sum_{\substack{\omega\in\Omega \\ (X(\omega)-\mathsf{E}(X))^2\geq\alpha}} (X(\omega)-\mathsf{E}(X))^2\mathsf{Pr}(\omega) \\
&\geq \sum_{\substack{\omega\in\Omega \\ (X(\omega)-\mathsf{E}(X))^2\geq\alpha}} \alpha\mathsf{Pr}(\omega) \\
&= \alpha\mathsf{Pr}\big((X-\mathsf{E}(X))^2\geq\alpha\big),
\end{aligned}$$

which yields the desired inequality. □

The French know this inequality as the *Bienaymé–Chebyshev's Inequality*. Bienaymé proved this inequality in 1853, before Chebyshev who published it in 1867. However, it was Chebyshev who recognized its significance.[2] Note that Chebyshev's Inequality can also be stated as

$$\mathsf{Pr}\big(|X-\mathsf{E}(X)|\geq\alpha\big)\leq\frac{\mathsf{Var}(X)}{\alpha^2}.$$

It is also convenient to restate the Chebyshev's Inequality in terms of the standard deviation $\sigma=\sqrt{\mathsf{Var}(X)}$ of X, to write $\mathsf{E}(X)=\mu$, and to replace α by $c^2\mathsf{Var}(X)$, and we get: For every $c>0$,

$$\mathsf{Pr}(|X-\mu|\geq c\sigma)\leq\frac{1}{c^2};$$

equivalently

$$\mathsf{Pr}(|X-\mu|<c\sigma)\geq 1-\frac{1}{c^2}.$$

This last inequality says that a random variable will lie within $c\sigma$ of its mean with probability at least $1-1/c^2$. If $c=10$, the random variable will lie between $\mu-10\sigma$ and $\mu+10\sigma$ at least 99% of the time.

We can apply the Chebyshev Inequality to the experiment where we roll two fair dice. We found that $\mu=7$ and $\sigma^2=35/6$ (for one roll). If we assume that we

[2] Still, Bienaymé is well loved!

perform n independent trials, then the total value of the n rolls has expecation $7n$ and the variance if $35n/6$. It follows that the sum will be between

$$7n - 10\sqrt{\frac{35n}{6}} \quad \text{and} \quad 7n + 10\sqrt{\frac{35n}{6}}$$

at least 99% of the time. If $n = 10^6$ (a million rolls), then the total value will be between 6.976 million and 7.024 million more than 99% of the time.

Another interesting consequence of the Chebyshev's Inequality is this. Suppose we have a random variable X on some discrete probability space (Ω, Pr). For any n, we can form the product space (Ω^n, Pr) as explained in Definition 6.6, with

$$\mathsf{Pr}(\omega_1, \ldots, \omega_n) = \mathsf{Pr}(\omega_1) \cdots \mathsf{Pr}(\omega_n), \quad \omega_i \in \Omega, i = 1, \ldots, n.$$

Then, we define the random variable X_k on the product space by

$$X_k(\omega_1, \ldots, \omega_n) = X(\omega_k).$$

It is easy to see that the X_k are independent. Consider the random variable

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

We can think of S as taking n independent "samples" from Ω and adding them together. By our previous discussion, S has mean $n\mu$ and standard deviation $\sigma\sqrt{n}$, where μ is the mean of X and σ is its standard deviation. The Chebyshev's Inequality implies that the average

$$\frac{X_1 + \cdots + X_n}{n}$$

will lie between $\sigma - 10\sigma/\sqrt{n}$ and $\sigma + 10\sigma/\sqrt{n}$ at least 99% of the time. This implies that if we choose n large enough, then the average of n samples will almost always be very near the expected value $\mu = \mathsf{E}(X)$.

This concludes our elementary introduction to discrete probability. The reader should now be well prepared to move on to Grimmett and Stirzaker [6] or Venkatesh [14]. Among the references listed at the end of this chapter, let us mention the classical volumes by Feller [3, 4], and Shiryaev [13].

The next three sections are devoted to more advanced topics and are optional.

6.6 Limit Theorems; A Glimpse

The behavior of the average sum of n independent samples described at the end of Section 6.5 is an example of a *weak law of large numbers*. A precise formulation of such a result is shown below. A version of this result was first shown by Jacob Bernoulli and was published by his nephew Nicholas in 1713. Bernoulli did not have

Chebyshev's Inequality at this disposal (since Chebyshev Inequality was proved in 1867), and he had to resort to a very ingenious proof.

Fig. 6.10 Jacob (Jacques) Bernoulli (1654–1705)

Theorem 6.1. *(Weak Law of Large Numbers ("Bernoulli's Theorem")) Let* $X_1, X_2, \ldots, X_n, \ldots$ *be a sequence of random variables. Assume that they are independent, that they all have the same distribution, and let* μ *be their common mean and* σ^2 *be their common variance (we assume that both exist). Then, for every* $\varepsilon > 0$*,*

$$\lim_{n \mapsto \infty} \mathsf{Pr}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) = 0.$$

Proof. As earlier,

$$\mathsf{E}\left(\frac{X_1 + \cdots + X_n}{n} \right) = \mu$$

and because the X_i are independent,

$$\mathsf{Var}\left(\frac{X_1 + \cdots + X_n}{n} \right) = \frac{\sigma^2}{n}.$$

Then, we apply Chebyshev's Inequality and we obtain

$$\mathsf{Pr}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{\sigma^2}{n\varepsilon^2},$$

which proves the result. □

The locution *independent and identically distributed* random variables is often used to say that some random variables are independent and have the same distribution. This locution is abbreviated as *i.i.d.* Probability books are replete with i.i.d.'s

Another remarkable limit theorem has to do with the limit of the distribution of the random variable

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}},$$

where the X_i are i.i.d random variables with mean μ and variance σ. Observe that the mean of $X_1 + \cdots + X_n$ is $n\mu$ and its variance is $\sigma\sqrt{n}$, since the X_i are assumed to be i.i.d.

We have not discussed a famous distribution, the normal or Gaussian distribution, only because it is a continuous distribution. The *standard normal distribution* is the cumulative distribution function Φ whose density function is given by

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

that is,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy.$$

The function $f(x)$ decays to zero very quickly and its graph has a bell–shape. More generally, we say that a random variable X is *normally distributed with parameters* μ *and* σ^2 (and that X has a *normal distribution*) if its density function is the function

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

Figure 6.11 shows some examples of normal distributions.

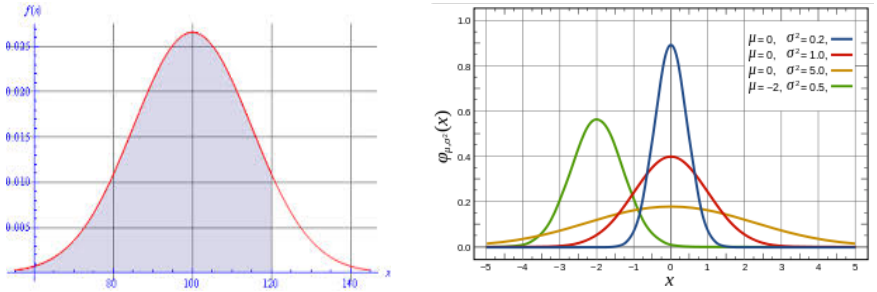

Fig. 6.11 Examples of normal distributions

Using a little bit of calculus, it is not hard to show that if a random variable X is normally distributed with parameters μ and σ^2, then its mean and variance are given by

$$\mathsf{E}(X) = \mu,$$
$$\mathsf{Var}(X) = \sigma^2.$$

The normal distribution with parameters μ and σ^2 is often denoted by $\mathcal{N}(\mu, \sigma^2)$. The standard case corresponds to $\mu = 0$ and $\sigma = 1$.

The following theorem was first proved by de Moivre in 1733 and generalized by Laplace in 1812. De Moivre introduced the normal distribution in 1733. However, it was Gauss who showed in 1809 how important the normal distribution (alternatively Gaussian distribution) really is.

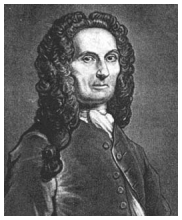

Fig. 6.12 Abraham de Moivre (1667–1754) (left), Pierre–Simon Laplace (1749–1827) (middle), Johann Carl Friedrich Gauss (1777–1855) (right)

Theorem 6.2. *(de Moivre–Laplace Limit Theorem) Consider n repeated independent Bernoulli trials (coin flips) X_i, where the probability of success is p. Then, for all $a < b$,*

$$\lim_{n\mapsto\infty} \mathsf{Pr}\left(a \leq \frac{X_1+\cdots+X_n-np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b)-\Phi(a).$$

Observe that now, we have two approximations for the distribution of a random variable $X = X_1+\cdots+X_n$ with a binomial distribution. When n is large and p is small, we have the Poisson approximation. When $np(1-p)$ is large, the normal approximation can be shown to be quite good.

Theorem 6.2 is a special case of the following important theorem known as *central limit theorem*.

Theorem 6.3. *(Central Limit Theorem) Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables. Assume that they are independent, that they all have the same distribution, and let μ be their common mean and σ^2 be their common variance (we assume that both exist). Then, the distribution of the random variable*

$$\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution as n goes to infinity. This means that for every real a,

$$\lim_{n\mapsto\infty} \mathsf{Pr}\left(\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{1}{2}x^2}.$$

We lack the machinery to prove this theorem. This machinery involves characteristic functions and various limit theorems. We refer the interested reader to Ross [11]

(Chapter 8), Grimmett and Stirzaker [6] (Chapter 5), Venkatesh [14], and Shiryaev [13] (Chapter III).

The central limit theorem was originally stated and proved by Laplace but Laplace's proof was not entirely rigorous. Laplace expanded a great deal of efforts in estimating sums of the form

$$\sum_{\substack{k \\ k \leq np + x\sqrt{np(1-p)}}} \binom{n}{k} p^k (1-p)^{n-k}$$

using Stirling's formula.

Reading Laplace's classical treatise [7, 8] is an amazing experience. The introduction to Volume I is 164 pages long! Among other things, it contains some interesting philosophical remarks about the role of probability theory, for example on the reliability of the testimony of witnesses. It is definitely worth reading. The second part of Volume I is devoted to the theory of generating functions, and Volume II to probability theory proper. Laplace's treatise was written before 1812, and even though the factorial notation was introduced in 1808, Laplace does not use it, which makes for complicated expressions. The exposition is clear, but it is difficult to read this treatise because definitions and theorems are not clearly delineated. A version of the central limit theorem is proved in Volume II, Chapter III; page 306 contains a key formula involving the Gaussian distribution, although Laplace does not refer to it by any name (not even as normal distribution). Anybody will be struck by the elegance and beauty of the typesetting. Lyapunov gave the first rigorous proof of the central limit theorem around 1901.

Fig. 6.13 Pierre–Simon Laplace (1749–1827) (left), Aleksandr Mikhailovich Lyapunov (1857-1918) (right)

The following example from Ross [11] illustrates how the central limit theorem can be used.

Example 6.29. An astronomer is interested in measuring the distance, in light-years, from his observatory to a distant star. Although the astronomer has a measuring technique, he knows that, because of changing atmospheric conditions and normal error, each time a measurement is made it will not be the exact distance, but merely an approximation. As a result, the astronomer plans to make a series of measurements

and then use the average value of these measurements as his estimated value of the actual distance.

If the astronomer believes that the values of the measurements are independent and identically distributed random variables having a common mean d and a common variance 4 (light-years), how many measurements need he make to be reasonably sure that his estimated distance is accurrate to within ± 0.5 light-years?

Suppose that the astronomer makes n observations, and let $X_1, \ldots, X_n$ be the n measurements. By the central limit theorem, the random variable

$$Z_n = \frac{X_1 + \cdots + X_n - nd}{2\sqrt{n}}$$

has approximately a normal distribution. Hence,

$$\begin{aligned} \mathsf{Pr}\left(-\frac{1}{2} \le \frac{X_1 + \cdots + X_n}{n} \le \frac{1}{2}\right) &= \mathsf{Pr}\left(-\frac{1}{2}\frac{\sqrt{n}}{2} \le Z_n \le \frac{1}{2}\frac{\sqrt{n}}{2}\right) \\ &\approx \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) \\ &= 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1. \end{aligned}$$

If the astronomer wants to be 95% certain that his estimated value is accurrate to within 0.5 light year, he should make n^* measurements, where n^* is given by

$$2\Phi\left(\frac{\sqrt{n^*}}{4}\right) - 1 = 0.95,$$

that is,

$$\Phi\left(\frac{\sqrt{n^*}}{4}\right) = 0.975.$$

Using tables for the values of the function Φ, we find that

$$\frac{\sqrt{n^*}}{4} = 1.96,$$

which yields

$$n^* \approx 61.47.$$

Since n should be an integer, the astronomer should make 62 observations.

The above analysis relies on the assumption that the distribution of Z_n is well approximated by the normal distribution. If we are concerned about this point, we can use Chebyshev's inequality. If we write

$$S_n = \frac{X_1 + \cdots + X_n - nd}{2},$$

we have

$$\mathsf{E}(S_n) = d \quad \text{and} \quad \mathsf{Var}(S_n) = \frac{4}{n},$$

so by Chebyshev's inequality, we have

$$\mathsf{Pr}\left(|S_n - d| > \frac{1}{2}\right) \leq \frac{4}{n(1/2)^2} = \frac{16}{n}.$$

Hence, if we make $n = 16/0.05 = 320$ observations, we are 95% certain that the estimate will be accurate to within 0.5 light year.

The method of making repeated measurements in order to "average" errors is applicable to many different situations (geodesy, astronomy, etc.).

There are generalizations of the central limit theorem to independent but not necessarily identically distributed random variables. Again, the reader is referred to Ross [11] (Chapter 8), Grimmett and Stirzaker [6] (Chapter 5), and Shiryaev [13] (Chapter III).

There is also the famous *strong law of large numbers* due to Andrey Kolmogorov proved in 1933 (with an earlier version proved in 1909 by Émile Borel). In order to state the strong law of large numbers, it is convenient to define various notions of convergence for random variables.

Fig. 6.14 Félix Edouard Justin Émile Borel (1871–1956) (left), Andrey Nikolaevich Kolmogorov (1903–1987) (right)

Definition 6.13. Given a sequence of random variable $X_1, X_2, \ldots, X_n, \ldots$, and some random variable X (on the same probability space (Ω, Pr)), we have the following definitions:

1. We say that X_n *converges to* X *almost surely* (abbreviated *a.s.*), denoted by $X_n \xrightarrow{\text{a.s.}} X$, if
$$\mathsf{Pr}(\{\omega \in \Omega \mid \lim_{n \mapsto \infty} X_n(\omega) = X(\omega)\}) = 1.$$
2. We say that X_n *converges to* X *in rth mean*, with $r \geq 1$, denoted $X_n \xrightarrow{r} X$, if $\mathsf{E}(|X_n^r|)$ is finite for all n and if
$$\lim_{n \mapsto \infty} \mathsf{E}\big(|X_n - X|^r\big) = 0.$$

3. We say that X_n *converges to* X *in probability*, denoted $X_n \xrightarrow{\text{P}} X$, if for every $\varepsilon > 0$,
$$\lim_{n\mapsto\infty} \mathsf{Pr}(|X_n - X| > \varepsilon) = 0.$$

4. We say that X_n *converges to* X *in distribution*, denoted $X_n \xrightarrow{\text{D}} X$, if
$$\lim_{n\mapsto\infty} \mathsf{Pr}(X_n \leq x) = \mathsf{Pr}(X \leq x),$$
for every $x \in \mathbb{R}$ for which $F(x) = \mathsf{Pr}(X \leq x)$ is continuous.

Convergence of type (1) is also called convergence *almost everywhere* or *convergence with probability* 1. Almost sure convergence can be stated as the fact that the set
$$\{\omega \in \Omega \mid X_n(\omega) \text{ does not converge to } X(\omega)\}$$
of outcomes for which convergence fails has probability 0.

It can be shown that both convergence almost surely and convergence in rth mean imply convergence in probability, which implies convergence in distribution. All converses are false. Neither convergence almost surely nor convergence in rth mean imply the other. For proofs, Interested readers should consult Grimmett and Stirzaker [6] (Chapter 7) and Shiryaev [13] (Chapter III).

Observe that the convergence of the weak law of large numbers is convergence in probability, and the convergence of the central limit theorem is convergence in distribution.

The following beautiful result was obtained by Kolmogorov (1933).

Theorem 6.4. *(Strong Law of Large Numbers, Kolmogorov) Let* $X_1, X_2, \ldots, X_n, \ldots$ *be a sequence of random variables. Assume that they are independent, that they all have the same distribution, and let* μ *be their common mean and* $\mathsf{E}(X_1^2)$ *be their common second moment (we assume that both exist). Then,*
$$\frac{X_1 + \cdots + X_n}{n}$$
converges almost surely and in mean square to $\mu = \mathsf{E}(X_1)$.

The proof is beyond the scope of this book. Interested readers should consult Grimmett and Stirzaker [6] (Chapter 7), Venkatesh [14], and Shiryaev [13] (Chapter III). Fairly accessible proofs under the additional assumption that $\mathsf{E}(X_1^4)$ exists can be found in Brémaud [2], and Ross [11].

Actually, for almost sure convergence, the assumption that $\mathsf{E}(X_1^2)$ exists is redundant provided that $\mathsf{E}(|X_1|)$ exists, in which case $\mu = \mathsf{E}(|X_1|)$, but the proof takes some work; see Brémaud [2] (Chapter 1, Section 8.4) and Grimmett and Stirzaker [6] (Chapter 7). There are generalizations of the strong law of large numbers where the independence assumption on the X_n is relaxed, but again, this is beyond the scope of this book.

6.7 Generating Functions; A Glimpse

If a random variables X on some discrete probability space (Ω, Pr) takes nonnegative integer values, then we can define its *probability generating function* (for short *pgf*) $G_X(z)$ as

$$G_X(z) = \sum_{k\geq 0} \mathsf{Pr}(X = k) z^k,$$

which can also be expressed as

$$G_X(z) = \sum_{\omega\in\Omega} \mathsf{Pr}(\omega) z^{X(\omega)} = \mathsf{E}(z^X).$$

Therefore

$$G_X(z) = \mathsf{E}(z^X).$$

Note that

$$G_X(1) = \sum_{\omega\in\Omega} \mathsf{Pr}(\omega) = 1,$$

so the radius of convergence of the power series $G_X(z)$ is at least 1. The nicest property about pgf's is that they usually simplify the computation of the mean and variance. For example, we have

$$\begin{aligned}
\mathsf{E}(X) &= \sum_{k\geq 0} k\mathsf{Pr}(X = k) \\
&= \sum_{k\geq 0} \mathsf{Pr}(X = k) \cdot kz^{k-1}\big|_{z=1} \\
&= G'_X(1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathsf{E}(X^2) &= \sum_{k\geq 0} k^2\mathsf{Pr}(X = k) \\
&= \sum_{k\geq 0} \mathsf{Pr}(X = k) \cdot (k(k-1)z^{k-2} + kz^{k-1})\big|_{z=1} \\
&= G''_X(1) + G'_X(1).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathsf{E}(X) &= G'_X(1) \\
\mathsf{Var}(X) &= G''_X(1) + G'_X(1) - (G'_1(1))^2.
\end{aligned}$$

Remark: The above results assume that $G'_X(1)$ and $G''_X(1)$ are well defined, which is the case if the radius of convergence of the power series $G_X(z)$ is greater than 1. If the radius of convergence of $G_X(z)$ is equal to 1 and if $\lim_{z\uparrow 1} G'_X(z)$ exists, then

$$\mathsf{E}(X) = \lim_{z\uparrow 1} G'_X(z),$$

and similarly if $\lim_{z\uparrow 1} G''_X(z)$ exists, then

$$\mathsf{E}(X^2) = \lim_{z\uparrow 1} G''_X(z).$$

The above facts follow from *Abel's theorem*, a result due to N. Abel. Abel's theorem

Fig. 6.15 Niels Henrik Abel (1802–1829)

states that if $G(x) = \sum_{n=0}^{\infty} a_n z^n$ is a real power series with radius of convergence $R = 1$ and if the sum $\sum_{n=0}^{\infty} a_n$ exists, which means that

$$\lim_{n\mapsto\infty} \sum_{i=0}^{n} a_i = a$$

for some $a \in \mathbb{R}$, then $G(z)$ can be extended to a uniformly convergent series on $[0,1]$ such that $\lim_{z\mapsto 1} G_X(z) = a$. For details, the reader is referred to Grimmett and Stirzaker [6] (Chapter 5) and Brémaud [2] (Appendix, Section 1.2).

However, as explained in Graham, Knuth and Patashnik [5], we may run into unexpected problems in using a closed form formula for $G_X(z)$. For example, if X is a random variable with the uniform distribution of order n , which means that X takes any value in $\{0,1,\ldots,n-1\}$ with equal probability $1/n$, then the pgf of X is

$$U_n = \frac{1}{n}(1+z+\cdots+z^{n-1}) = \frac{1-z^n}{n(1-z)}.$$

If we set $z = 1$ in the above closed-form expression, we get $0/0$. The computations of the derivatives $U'_X(1)$ and $U''_X(1)$ will also be problematic (although we can resort to L'Hospital's rule).

Fortunately, there is an easy fix. If $G(z) = \sum_{n\geq 0} a_n z^n$ is a power series that converges for some z with $|z| > 1$, then $G'(z) = \sum_{n\geq 0} n a_n z^{n-1}$ also has that property, and by Taylor's theorem, we can write

$$G(1+x) = G(1) + \frac{G'(1)}{1!}x + \frac{G''(1)}{2!}x^2 + \frac{G'''(1)}{3!}x^3 + \cdots.$$

It follows that all derivatives of $G(z)$ at $z = 1$ appear as coefficients when $G(1+x)$ is expanded in powers of x. For example, we have

$$\begin{aligned} U_n(1+x) &= \frac{(1+x)^n - 1}{nx} \\ &= \frac{1}{n}\binom{n}{1} + \frac{1}{n}\binom{n}{2}x + \frac{1}{n}\binom{n}{3}x^2 + \cdots + \frac{1}{n}\binom{n}{n}x^{n-1}. \end{aligned}$$

It follows that

$$U_n(1) = 1; \qquad U_n'(1) = \frac{n-1}{2}; \qquad U_n''(1) = \frac{(n-1)(n-2)}{3};$$

Then, we find that the mean is given by

$$\mu = \frac{n-1}{2}$$

and the variance by

$$\sigma^2 = U_n''(1) + U_n'(1) - (U_n'(1))^2 = \frac{n^2-1}{12}.$$

Another nice fact about pgf's is that the pdf of the sum $X+Y$ of two independent variables X and Y is the product their pgf's. This is because if X and Y are independent, then

$$\begin{aligned} \mathsf{Pr}(X+Y=n) &= \sum_{k=0}^{n} \mathsf{Pr}(X=k \text{ and } Y=n-k) \\ &= \sum_{k=0}^{n} \mathsf{Pr}(X=k)\mathsf{Pr}(Y=n-k), \end{aligned}$$

a convolution! Therefore, if X and Y are independent, then

$$G_{X+Y}(z) = G_X(z)G_Y(z).$$

If we flip a biased coin where the probability of tails is p, then the pgf for the number of heads after one flip is

$$H(z) = 1 - p + pz.$$

If we make n independent flips, then the pgf of the number of heads is

$$H(z)^n = (1-p+pz)^n.$$

This allows us to rederive the formulae for the mean and the variance. We get

$$\mu = (H^n(z)'(1) = nH'(1) = np,$$

and

$$\sigma^2 = n(H''(1) + H'(1) - (H'(1))^2) = n(0 + p - p^2) = np(1-p).$$

If we flip a biased coin repeatedly until heads first turns up, we saw that the random variable X that gives the number of trials n until the first occurrence of tails has the geometric distribution $f(n) = (1-p)^{n-1}p$. It follows that the pgf of X is

$$G_X(z) = pz + (1-p)pz^2 + \cdots + (1-p)^{n-1}pz^n + \cdots = \frac{pz}{1-(1-p)z}.$$

Since we are assuming that these trials are independent, the random variables that tell us that m heads are obtained has pgf

$$\begin{aligned} G_X(z) &= \left(\frac{pz}{1-(1-p)z}\right)^m \\ &= p^m z^m \sum_k \binom{m+k-1}{k} ((1-p)z)^k \\ &= \sum_j \binom{j-1}{j-m} p^m (1-p)^{j-m} z^k. \end{aligned}$$

An an exercise, the reader should check that the pgf of a Poisson distribution with parameter λ is

$$G_X(z) = e^{\lambda(z-1)}.$$

More examples of the use of pgf can be found in Graham, Knuth and Patashnik [5].

Another interesting generating function is the *moment generating function* $M_X(t)$. It is defined as follows: for any $t \in \mathbb{R}$,

$$M_X(t) = \mathsf{E}(e^{tX}) = \sum_x e^{tx} f(x),$$

where $f(x)$ is the mass function of X. If X is a continuous random variable with density function f, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

The main problem with the moment generating function is that it is not always defined for all $t \in \mathbb{R}$. If $M_X(t)$ converges absolutely on some open interval $(-r, r)$ with $r > 0$, then its nth derivative for $t = 0$ is given by

$$M^{(n)}(0) = \sum_x x^n e^{tx} f(x)\big|_{t=0} = \sum_x x^n f(x) = \mathsf{E}(X^n).$$

Therefore, the moments of X are all defined and given by

$$\mathsf{E}(X^n) = M^{(n)}(0).$$

Within the radius of convergence of $M_X(t)$, we have the Taylor expansion

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\mathsf{E}(X^k)}{k!} t^k.$$

This shows that $M_X(t)$ is the *exponential generating function* of the sequence of moments $(\mathsf{E}(X^n))$; see Graham, Knuth and Patashnik [5]. If X is a continuous random variable, then the function $M_X(-t)$ is the *Laplace transform* of the density function f.

Furthermore, if X and Y are independent, then $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$, so we have

$$\mathsf{E}\big((X+Y)^n\big) = \sum_{k=0}^{n} \binom{n}{k} \mathsf{E}(X^k Y^{n-k}) = \sum_{k=0}^{n} \binom{n}{k} \mathsf{E}(X)^k \mathsf{E}(Y)^{n-k},$$

and since

$$\begin{aligned} M_{X+Y}(t) &= \sum_n \frac{\mathsf{E}\big((X+Y)^n\big)}{n!} t^n \\ &= \frac{1}{n!} \left(\sum_{k=0}^{n} \binom{n}{k} \mathsf{E}(X)^k \mathsf{E}(Y)^{n-k} \right) t^n \\ &= \sum_n \frac{\mathsf{E}(X)^k}{k!} \frac{\mathsf{E}(Y)^{n-k}}{(n-k)!} t^n \\ &= \sum_n \frac{\mathsf{E}(X^k)}{k!} \frac{\mathsf{E}(Y^{n-k})}{(n-k)!} t^n. \end{aligned}$$

But, this last term is the coefficient of t^n in $M_X(t)M_Y(t)$. Therefore, as in the case of pgf's, if X and Y are independent, then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Another way to prove the above equation is to use the fact that if X and Y are independent random variables, then so are e^{tX} and e^{tY} for any fixed real t. Then,

$$\mathsf{E}(e^{t(X+Y)}) = \mathsf{E}(e^{tX}e^{tY}) = \mathsf{E}(e^{tX})\mathsf{E}(e^{tY}).$$

Remark: If the random variable X takes nonnegative integer values, then it is easy to see that

$$M_X(t) = G_X(e^t),$$

where G_X is the generating function of X, so M_X is defined over some open interval $(-r,r)$ with $r > 0$ and $M_X(t) > 0$ on this interval. Then, the function $K_X(t) = \ln M_X(t)$ is well defined, and it has a Taylor expansion

$$K_X(t) = \frac{\kappa_1}{1!}t + \frac{\kappa_2}{2!}t^2 + \frac{\kappa_3}{3!}t^3 + \cdots + \frac{\kappa_n}{n!}t^n + \cdots. \qquad (*)$$

The numbers κ_n are called the *cumulants* of X. Since

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\mu_n}{n!} t^n,$$

where $\mu_n = \mathsf{E}(E^n)$ is the nth moment of X, by taking exponentials on both sides of $(*)$, we get relations between the cumulants and the moments, namely:

$$\begin{aligned}
\kappa_1 &= \mu_1 \\
\kappa_2 &= \mu_2 - \mu_1^2 \\
\kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \\
\kappa_4 &= \mu_4 - 4\mu_1\mu_4 + 12\mu_1^2\mu_2 - 3\mu_2^2 - 6\mu_1^4 \\
&\vdots
\end{aligned}$$

Notice that κ_1 is the mean and κ_2 is the variance of X. Thus, it appears that the cumulants are the natural generalization of the mean and variance. Furthermore, because logs are taken, all cumulants of the sum of two independent random variables are additive, just as the mean and variance. This property makes cumulants more important than moments.

The third generating function associtaed with a random variable X, and the most important, is the *characteristic function* $\varphi_X(t)$, defined by

$$\varphi_X(t) = \mathsf{E}(e^{itX}) = \mathsf{E}(\cos tX) + i\mathsf{E}(\sin tX),$$

for all $t \in \mathbb{R}$. If f is the mass function of X, we have

$$\varphi_X(t) = \sum_x e^{itx} f(x) = \sum_x \cos(tx) f(x) + i \sum_x \sin(tx) f(x),$$

a complex function of the real variable t. The "innocent" insertion of i in the exponent has the effect that $|e^{itX}| = 1$, so $\varphi_X(t)$ is defined for all $t \in \mathbb{R}$.

If X is a continuous random variable with density function f, then

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

Up to sign and to a change of variable, $\varphi_X(t)$ is basically the Fourier transform of f. Traditionally the *Fourier transform* $\widehat{f}$ of f is given by

$$\widehat{f}(t) = \int_{-\infty}^{\infty} e^{-2\pi itx} f(x) dx.$$

Next, we summarize some of the most important properties of φ_X without proofs. Details can be found in Grimmett and Stirzaker [6] (Chapter 5).

The characteristic function φ_X of a random variable satisfies the following properties:

1. $\varphi_X(0) = 1$, $|\varphi_X(t)| \leq 1$.

2. φ_X is uniformly continuous on $\mathbb{R}$.
3. If $\varphi^{(n)}$ exists, then $\mathsf{E}(|E^k|)$ is finite if k is even, and $\mathsf{E}(|E^{k-1}|)$ is finite if k is odd.
4. If X and Y are independent, then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

   The proof is essentially the same as the one we gave for the moment generating function, modulo powers of i.
5. If X is a random variable, for any two reals a, b,

$$\varphi_{aX+b}(t) = e^{itb}\varphi_X(at).$$

Given two random variables X and Y, their *joint characteristic function* $\varphi_{X,Y}(x,y)$ is defined by

$$\varphi_{X,Y}(x,y) = \mathsf{E}(e^{ixX}e^{iyY}).$$

Then, X and Y are independent iff

$$\varphi_{X,Y}(x,y) = \varphi_X(x)\varphi_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

In general, if all the moments $\mu_n = \mathsf{E}(X^n)$ of a random variable X are defined, these moments do not uniquely define the distibution F of X. There are examples of distinct distributions F (for X) and G (for Y) such that $\mathsf{E}(X^n) = \mathsf{E}(Y^n)$ for all n; see Grimmett and Stirzaker [6] (Chapter 5).

However, if the moment generating function of X is defined on some open interval $(-r, r)$ with $r > 0$, then $M_X(t)$ defines the distribution F of X uniquely.

The reason is that in this case, the characteristic function φ_X is holomorphic on the strip $|\mathrm{Im}(z)| < r$, and then M_X can extended to that strip to a holomorphic function such that $\varphi_X(t) = M_X(it)$. Furthermore, the characteristic function φ_X determines the distribution F of X uniquely. This is a rather deep result which is basically a version of Fourier inversion. If X is a continuous random variable with density function f, then

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt,$$

for every x for which f is differentiable.

If the distribution F is not given as above, it is still possible to prove the following result (see Grimmett and Stirzaker [6] (Chapter 5)):

Theorem 6.5. *Two random variables X and Y have the same characteristic function iff they have the same distribution.*

As a corollary, if the moment generating functions M_X and M_Y are defined on some interval $(-r, r)$ with $r > 0$ and if $M_X = M_Y$, then X and Y have the same distribution. In computer science, this condition seems to be always satisfied.

If X is a discrete random variable that takes integer values, then

$$f(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-itk}\varphi_X(t)dt;$$

see Grimmett and Stirzaker [6] (Chapter 5, Exercise 4).

There are also some useful continuity theorems which can be found in Grimmett and Stirzaker [6] (Chapter 5). In the next section, we use the moment generating function to obtain bounds on tail distributions.

6.8 Chernoff Bounds

Given a random variable X, it is often desirable to have information about probabilities of the form $\mathsf{Pr}(X \geq a)$ (for some real a). In particular, it may be useful to know how quickly such a probability goes to zero as a becomes large (in absolute value). Such probabilities are called *tail distributions*. It turns out that the moment generating function M_X (if it exists) yields some useful bounds by applying a very simple inequality to M_X known as *Markov's inequality* and due to the mathematician Andrei Markov, a major contributor to probability theory (the inventor of Markov chains).

Fig. 6.16 Andrei Andreyevich Markov (1856–1922)

Proposition 6.17. *(Markov's Inequality) Let X be a random variable and assume that X is nonnegative. Then, for every $a > 0$, we have*

$$\mathsf{Pr}(X \geq a) \leq \frac{\mathsf{E}(X)}{a}.$$

Proof. Let I_a be the random variable defined so that

$$I_a = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Since $X \geq 0$, we have

$$I_a \leq \frac{X}{a}. \qquad (*)$$

Also, since I_a takes only the values 0 and 1, $\mathsf{E}(I_a) = \mathsf{Pr}(X \geq a)$. By taking expectations in $(*)$, we get

$$\mathsf{E}(I_a) \leq \frac{\mathsf{E}(X)}{a},$$

which is the desired inequality since $\mathsf{E}(I_a) = \mathsf{Pr}(X \geq a)$. □

If we apply Markov's inequality to the moment generating function $M_X = \mathsf{E}(E^{tX})$ we obtain exponential bounds known as *Chernoff bounds*, after Herman Chernoff.

Fig. 6.17 Herman Chernoff (1923–)

Proposition 6.18. *(Chernoff Bounds) Let X be a random variable and assume that the moment generating function $M_X = \mathsf{E}(e^{tX})$ is defined. Then, for every $a > 0$, we have*

$$\mathsf{Pr}(X \geq a) \leq \min_{t>0} e^{-ta} M_X(t)$$
$$\mathsf{Pr}(X \leq a) \leq \min_{t<0} e^{-ta} M_X(t).$$

Proof. If $t > 0$, by Markov's inequality applied to $M_X(t) = \mathsf{E}(e^{tX})$, we get

$$\begin{aligned} \mathsf{Pr}(X \geq a) &= \mathsf{Pr}(e^{tX} \geq e^{ta}) \\ &\leq \mathsf{E}(e^{tX}) e^{-ta}, \end{aligned}$$

and if $t < 0$, we get

$$\begin{aligned} \mathsf{Pr}(X \leq a) &= \mathsf{Pr}(e^{tX} \leq e^{ta}) \\ &\leq \mathsf{E}(e^{tX}) e^{-ta}, \end{aligned}$$

which imply both inequalities of the proposition. □

In order to make good use of the Chernoff bounds, one needs to find for which values of t the function $e^{-ta} M_X(t)$ is minimum. Let us give a few examples.

Example 6.30. If X has a standard normal distribution, then it is not hard to show that

$$M(t) = e^{t^2/2}.$$

Consequently, for any $a > 0$ and all $t > 0$, we get

$$\mathsf{Pr}(X \geq a) \leq e^{-ta} e^{t^2/2}.$$

The value t that minimizes $e^{t^2/2-ta}$ is the value that minimizes $t^2/2 - ta$, namely $t = a$. Thus, for $a > 0$, we have

$$\mathsf{Pr}(X \geq a) \leq e^{-a^2/2}.$$

Similarly, for $a < 0$, we obtain

$$\mathsf{Pr}(X \leq a) \leq e^{-a^2/2}.$$

The function on the right hand side decays to zero very quickly.

Example 6.31. Let us now consider a random variable X with a Poisson distribution with parameter λ. It is not hard to show that

$$M(t) = e^{\lambda(e^t - 1)}.$$

Applying the Chernoff bound, for any nonnegative integer k and all $t > 0$, we get

$$\mathsf{Pr}(X \geq k) \leq e^{\lambda(e^t - 1)} e^{-kt}.$$

Using calculus, we can show that the function on the right hand side has a minimum when $\lambda(e^t - 1) - kt$ is minimum, and this is when $e^t = k/\lambda$. If $k > \lambda$ and if we let $e^t = k/\lambda$ in the Chernoff bound, we obtain

$$\mathsf{Pr}(X \geq k) \leq e^{\lambda(k/\lambda - 1)} \left(\frac{\lambda}{k}\right)^k,$$

which is equivalent to

$$\mathsf{Pr}(X \geq k) \leq \frac{e^{-\lambda}(e\lambda)^k}{k^k}.$$

Our third example is taken from Mitzenmacher and Upfal [10] (Chapter 4).

Example 6.32. Suppose we have a sequence of n random variables $X_1, X_2, \ldots, X_n$, such that each X_i is a Bernoulli variable (with value 0 or 1) with probability of success p_i, and assume that these variables are independent. Such sequences are often called *Poisson trials*. We wish to apply the Chernoff bounds to the random variable

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

We have

$$\mu = \mathsf{E}(X) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} p_i.$$

The moment generating function of X_i is given by

$$\begin{aligned} M_{X_i}(t) &= \mathsf{E}(e^{tX_i}) \\ &= p_i e^t + (1-p_i) \\ &= 1 + p_i(e^t - 1). \end{aligned}$$

Using the fact that $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we obtain the bound

$$M_{X_i}(t) \leq e^{p_i(e^t-1)}.$$

Since the X_i are independent, we know from Section 6.7 that

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} \\ &= e^{\sum_{i=1}^{n} p_i(e^t-1)} \\ &= e^{\mu(e^t-1)}. \end{aligned}$$

Therefore,

$$M_X(t) \leq e^{\mu(e^t-1)} \quad \text{for all } t.$$

The next step is to apply the Chernoff bounds. Using a little bit of calculus, we obtain the following result proved in Mitzenmacher and Upfal [10] (Chapter 4).

Proposition 6.19. *Given n independent Bernoulli variables $X_1, \ldots, X_n$ with success probability p_i, if we let $\mu = \sum_{i=1}^{n} p_i$ and $X = X_1 + \cdots + X_n$, then for any δ such that $0 < \delta < 1$, we have*

$$\mathsf{Pr}(|X - \mu| \geq \delta\mu) \leq 2e^{-\frac{\mu\delta^2}{3}}.$$

An an application, if the X_i are independent flips of a fair coin ($p_i = 1/2$), then $\mu = n/2$, and by picking $\delta = \frac{6\ln n}{n}$, it is easy to show that

$$\mathsf{Pr}\left(\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{6n\ln n}\right) \leq 2e^{-\frac{\mu\delta^2}{3}} = \frac{2}{n}.$$

This shows that the concentrations of the number of heads around the mean $n/2$ is very tight. Most of the time, the deviations from the mean are of the order $O(\sqrt{n\ln n})$. Another simple calculation using the Chernoff bounds shows that

$$\mathsf{Pr}\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq 2e^{-\frac{n}{24}}.$$

This is a much better bound than the bound provided by the Chebyshev inequality:

$$\mathsf{Pr}\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}.$$

Ross [11] and Mitzenmacher and Upfal [10] consider the situation where a gambler is equally likely to win or lose one unit on every play. Assuming that these random variables X_i are independent, and that

$$\mathsf{Pr}(X_i = 1) = \mathsf{Pr}(X_i = -1) = \frac{1}{2},$$

let $S_n = \sum_{i=1}^{n} X_i$ be the gamblers's winning after n plays. It is easy that to see that the moment generating function of X_i is

$$M_{X_i}(t) = \frac{e^t + e^{-t}}{2}.$$

Using a little bit of calculus, one finds that

$$M_{X_i}(t) \leq e^{\frac{t^2}{2}}.$$

Since the X_i are independent, we obtain

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = (M_{X_i}(t))^n \leq e^{\frac{nt^2}{2}}, \quad t > 0.$$

The Chernoff bound yields

$$\mathsf{Pr}(S_n \geq a) \leq e^{\frac{nt^2}{2} - ta}, \quad t > 0.$$

The minimum is achieved for $t = a/n$, and assuming that $a > 0$, we get

$$P(S_n \geq a) \leq e^{-\frac{a^2}{2n}}, \quad a > 0.$$

For example, if $a = 6$, we get

$$\mathsf{Pr}(S_{10} \geq 6) \leq e^{-\frac{36}{20}} \approx 0.1653.$$

We leave it as exercise to prove that

$$\mathsf{Pr}(S_n \geq 6) = \mathsf{Pr}(\text{gambler wins at least 8 of the first 10 games}) = \frac{56}{1024} \approx 0.0547.$$

Other examples of the use of Chernoff bounds can be found in Mitzenmacher and Upfal [10] and Ross [12]. There are also inequalities giving a lower bound on the probability $\mathsf{Pr}(X > 0)$, where X is a nonnegative random variable; see Ross [12] (Chapter 3), which contains other techniques to find bounds on probabilities, and the Poisson paradigm. Probabilistic methods also play a major role in Motwani and Raghavan [9].

6.9 Summary

This chapter provides an introduction to discrete probability theory. We define probability spaces (finite and countably infinite) and quickly get to random variables. We emphasize that random variables are more important that their underlying probability spaces. Notions such as expectation and variance help us to analyze the behavior of random variables even if their distributions are not known precisely. We give a number of examples of computations of expectations, including the coupon collector problem and a randomized version of quicksort.

The last three sections of this chapter contain more advanced material and are optional. The topics of these optional sections are generating functions (including the moment generating function and the characteristic function), the limit theorems (weak law of lage numbers, central limit theorem, and strong law of large numbers), and Chernoff bounds.

- We define: a finite *discrete probability space* (or finite *discrete sample space*), *outcomes* (or *elementary events*), and *events*.
- a *probability measure* (or *probability distribution*) on a sample space.
- a *discrete probability space*.
- a σ*-algebra*.
- *independent* events.
- We discuss the *birthday problem*.
- We give examples of *random variables*.
- We present a randomized version of the *quicksort* algorithm.
- We define: *random variables*, and their *probability mass functions* and *cumulative distribution functions*.
- *absolutely continuous* random variables and their *probability density functions*.
- We give examples of: the *binomial distribution*.
- the *geometric distribution*.
- We show how the *Poisson distribution* arises as the limit of a binomial distribution when n is large and p is small.
- We define a *conditional probability*.
- We present the "Monty Hall Problem."
- We introduce *probablity trees* (or *trees of possibilities*).
- We prove several of *Bayes' rules*.
- We define: the product of probability spaces.
- *independent* random variables.
- the *joint mass function* of two random variables, and the *marginal mass functions*.
- the *expectation* (or *expected value*, or *mean*) $\mathsf{E}(X) = \mu$ of a random variable X.
- We prove the *linearity* of expectation.
- We introduce *indicator functions* (*indicator variables*).
- We define functions of a random variables.
- We compute the expected value of the number of comparsions in the randomized version of quicksort.

- We define the *variance* $\mathsf{Var}(X)$ of a random variable X and the *standard deviation* σ of X by $\sigma = \sqrt{\mathsf{Var}(X)}$.
- We prove that $\mathsf{Var}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2$.
- We define the *moments* and the *central moments* of a random variable.
- We prove that if X and Y are uncorrelated random variables, then $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$; in particular, this equation holds if X and Y are independent.
- We prove: the *Cauchy-Schwarz inequality* for discrete random variables.
- *Cheybyshev's inequality* and give some of its applications.

The next three sections are optional.

- We state the *weak law of large numbers* (Bernoulli's theorem).
- We define the *normal distribution* (or *Gaussian distribution*).
- We sate the *central limit theorem* and present an application.
- We define various notions of convergence, including *almost sure convergence* and *convergence in probability*.
- We state Kolmogorov's *strong law of large numbers*.
- For a random variable that takes nonnegative integer values, we define the *probability generating function*, $G_X(z) = \mathsf{E}(z^X)$. We show how the derivatives of G_X at $z = 1$ can be used to compute the mean μ and the variance of X.
- If X and Y are independent random variables, then $G_{X+Y} = G_X G_Y$.
- We define the *moment generating function* $M_X(t) = \mathsf{E}(e^{tX})$ and show that $M_X^{(n)}(0) = \mathsf{E}(X^n)$.
- If X and Y are independent random variables, then $M_{X+Y} = M_X M_Y$.
- We define: the *cumulants* of X.
- the *characteristic function* $\varphi_X(t) = \mathsf{E}(e^{itX})$ of X and discuss some of its properties. Unlike the moment generating function, φ_X is defined for all $t \in \mathbb{R}$.
- If X and Y are independent random variables, then $\varphi_{X+Y} = \varphi_X \varphi_Y$. The distribution of a random variable is uniquely determined by its characteristic function.
- We prove: *Markov's inequality*.
- the general *Chernoff bounds* in terms of the moment generating function.
- We compute Chernoff bound for various distributions, including normal and Poisson.
- We obtain Chernoff bounds for *Poisson trials* (independent Bernoulli trials with success probablity p_i).

Problems

References

1. Joseph Bertrand. *Calcul des Probabilités*. New York, NY: Chelsea Publishing Company, third edition, 1907.
2. Pierre Brémaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulations, and Queues*. TAM No. 31. New York, NY: Springer, third edition, 2001.

3. William Feller. *An Introduction to Probability Theory and its Applications, Vol. 1* . New York, NY: Wiley, third edition, 1968.
4. William Feller. *An Introduction to Probability Theory and its Applications, Vol. 2* . New York, NY: Wiley, second edition, 1971.
5. Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation For Computer Science*. Reading, MA: Addison Wesley, second edition, 1994.
6. Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford, UK: Oxford University Press, third edition, 2001.
7. Pierre–Simon Laplace. *Théorie Analytique des Probabilités, Volume I*. Paris, France: Editions Jacques Gabay, third edition, 1820.
8. Pierre–Simon Laplace. *Théorie Analytique des Probabilités, Volume II*. Paris, France: Editions Jacques Gabay, third edition, 1820.
9. Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, first edition, 1995.
10. Michael Mitzenmacher and Eli Upfal. *Probability and Computing. Randomized Algorithms and Probabilistic Analysis*. Cambridge, UK: Cambridge University Press, first edition, 2005.
11. Sheldon Ross. *A First Course in Probability*. Upper Saddle River, NJ: Pearson Prentice Hall, eighth edition, 2010.
12. Sheldon Ross. *Probability Models for Computer Science*. San Diego, CA: Harcourt /Academic Press, first edition, 2002.
13. Albert Nikolaevich Shiryaev. *Probability*. GTM No. 95 . New York, NY: Springer, second edition, 1995.
14. Santosh S. Venkatesh. *The Theory of Probability: Explorations and Applications*. Cambridge, UK: Cambridge University Press, first edition, 2012.