A Primer on Probability

Victor Shoup

1 Basic definitions

Let Ω be a finite, non-empty set. A **probability distribution on** Ω is a function $\Pr : \Omega \to [0,1]$ that satisfies the following property:

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

The set Ω is called the **sample space of** $\Pr$

Intuitively, the elements of Ω represent the possible outcomes of a random experiment, where the probability of outcome $\omega \in \Omega$ is $\Pr(\omega)$.

For now, we restrict our discussion to finite sample spaces. Later, in §6, we generalize to *countably infinite* sample spaces.

Example 1. If we think of rolling a fair die, then setting $\Omega := \{1,2,3,4,5,6\}$, and $\Pr(\omega) := 1/6$ for all $\omega \in \Omega$, gives a probability distribution that naturally describes the possible outcomes of the experiment. □

Example 2. More generally, if Ω is any non-empty, finite set, and $\Pr(\omega) := 1/|\Omega|$ for all $\omega \in \Omega$, then $\Pr$ is called the **uniform distribution on** Ω. □

Example 3. A coin toss is an example of a **Bernoulli trial**, which in general is an experiment with only two possible outcomes: *success*, which occurs with probability p; and *failure*, which occurs with probability $q := 1 - p$. Of course, *success* and *failure* are arbitrary names, which can be changed as convenient. In the case of a coin, we might associate *success* with the outcome that the coin comes up *heads*. For a fair coin, we have $p = q = 1/2$; for a biased coin, we have $p \neq 1/2$. □

An **event** is a subset $\mathcal{A}$ of Ω, and the **probability of** $\mathcal{A}$ is defined to be

$$\Pr[\mathcal{A}] := \sum_{\omega \in \mathcal{A}} \Pr(\omega). \tag{2}$$

While an event is simply a subset of the sample space, when discussing the probability of an event (or other properties to be introduced later), the discussion always takes place relative to a particular probability distribution, which may be implicit from context.

For events $\mathcal{A}$ and $\mathcal{B}$, their union $\mathcal{A} \cup \mathcal{B}$ logically represents the event that *either* the event $\mathcal{A}$ *or* the event $\mathcal{B}$ occurs (or both), while their intersection $\mathcal{A} \cap \mathcal{B}$ logically represents the event that *both* $\mathcal{A}$ *and* $\mathcal{B}$ occur. For an event $\mathcal{A}$, we define its complement $\overline{\mathcal{A}} := \Omega \setminus \mathcal{A}$, which logically represents the event that $\mathcal{A}$ does *not* occur.

In working with events, one makes frequent use of the usual rules of Boolean logic. **De Morgan's law** says that for all events $\mathcal{A}$ and $\mathcal{B}$,

$$\overline{\mathcal{A} \cup \mathcal{B}} = \overline{\mathcal{A}} \cap \overline{\mathcal{B}} \quad \text{and} \quad \overline{\mathcal{A} \cap \mathcal{B}} = \overline{\mathcal{A}} \cup \overline{\mathcal{B}}.$$

We also have the **Boolean distributive law**: for all events $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$,

$$\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C}) \quad \text{and} \quad \mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C}).$$

Example 4. Continuing with Example 1, the event that the die has an odd value is $\mathcal{A} := \{1, 3, 5\}$, and we have $\Pr[\mathcal{A}] = 1/2$. The event that the die has a value greater than 2 is $\mathcal{B} := \{3, 4, 5, 6\}$, and $\Pr[\mathcal{B}] = 2/3$. The event that the die has a value that is at most 2 is $\overline{\mathcal{B}} = \{1, 2\}$, and $\Pr[\overline{\mathcal{B}}] = 1/3$. The event that the value of the die is odd *or* exceeds 2 is $\mathcal{A} \cup \mathcal{B} = \{1, 3, 4, 5, 6\}$, and $\Pr[\mathcal{A} \cup \mathcal{B}] = 5/6$. The event that the value of the die is odd *and* exceeds 2 is $\mathcal{A} \cap \mathcal{B} = \{3, 5\}$, and $\Pr[\mathcal{A} \cap \mathcal{B}] = 1/3$. □

Example 5. If Pr is the uniform distribution on a set Ω, and $\mathcal{A}$ is a subset of Ω, then $\Pr[\mathcal{A}] = |\mathcal{A}|/|\Omega|$. □

We next derive some elementary facts about probabilities of certain events, and relations among them. It is clear from the definitions that

$$\Pr[\emptyset] = 0 \quad \text{and} \quad \Pr[\Omega] = 1,$$

and that for every event $\mathcal{A}$, we have

$$\Pr[\overline{\mathcal{A}}] = 1 - \Pr[\mathcal{A}].$$

Now consider events $\mathcal{A}$ and $\mathcal{B}$, and their union $\mathcal{A} \cup \mathcal{B}$. We have

$$\Pr[\mathcal{A} \cup \mathcal{B}] \leq \Pr[\mathcal{A}] + \Pr[\mathcal{B}]; \tag{3}$$

moreover,

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] \quad \text{if } \mathcal{A} \text{ and } \mathcal{B} \text{ are disjoint}, \tag{4}$$

that is, if $\mathcal{A} \cap \mathcal{B} = \emptyset$. The exact formula for arbitrary events $\mathcal{A}$ and $\mathcal{B}$ is:

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cap \mathcal{B}]. \tag{5}$$

(3), (4), and (5) all follow from the observation that in the expression

$$\Pr[\mathcal{A}] + \Pr[\mathcal{B}] = \sum_{\omega \in \mathcal{A}} \Pr(\omega) + \sum_{\omega \in \mathcal{B}} \Pr(\omega),$$

the value $\Pr(\omega)$ is counted once for each $\omega \in \mathcal{A} \cup \mathcal{B}$, except for those $\omega \in \mathcal{A} \cap \mathcal{B}$, for which $\Pr(\omega)$ is counted twice.

Example 6. Alice rolls two dice, and asks Bob to guess a value that appears on either of the two dice (without looking). Let us model this situation by considering the uniform distribution on $\Omega := \{1, \ldots, 6\} \times \{1, \ldots, 6\}$, where for each pair $(s, t) \in \Omega$, s represents the value of the first die, and t the value of the second.

For $k = 1, \ldots, 6$, let $\mathcal{A}_k$ be the event that the first die is k, and $\mathcal{B}_k$ the event that the second die is k. Let $\mathcal{C}_k = \mathcal{A}_k \cup \mathcal{B}_k$ be the event that k appears on either of the two dice. For any fixed k, the event $\mathcal{A}_k$ consist of the 6 outcomes $(k, 1), \ldots, (k, 6)$, and so $\Pr[\mathcal{A}_k] = 6/36 = 1/6$; similarly, $\Pr[\mathcal{B}_k] = 1/6$; the event $\mathcal{A}_k \cap \mathcal{B}_k$ consists of the single outcome (k, k), and so $\Pr[\mathcal{A}_k \cap \mathcal{B}_k] = 1/36$.

So no matter what value k Bob chooses, the probability that this choice is correct is

$$\begin{aligned} \Pr[\mathcal{C}_k] &= \Pr[\mathcal{A}_k \cup \mathcal{B}_k] = \Pr[\mathcal{A}_k] + \Pr[\mathcal{B}_k] - \Pr[\mathcal{A}_k \cap \mathcal{B}_k] \\ &= 1/6 + 1/6 - 1/36 = 11/36, \end{aligned}$$

which is slightly less than the estimate $\Pr[\mathcal{A}_k] + \Pr[\mathcal{B}_k]$ obtained from (3). □

If $\{\mathcal{A}_i\}_{i\in I}$ is a family of events, indexed by some set I, we can naturally form the union $\bigcup_{i\in I}\mathcal{A}_i$ and intersection $\bigcap_{i\in I}\mathcal{A}_i$. If $I=\emptyset$, then by definition, the union is $\emptyset$, and by special convention, the intersection is the entire sample space Ω. Logically, the union represents the event that *some* $\mathcal{A}_i$ occurs, and the intersection represents the event that *all* the $\mathcal{A}_i$'s occur. De Morgan's law generalizes as follows:

$$\overline{\bigcup_{i\in I}\mathcal{A}_i}=\bigcap_{i\in I}\overline{\mathcal{A}}_i \quad\text{and}\quad \overline{\bigcap_{i\in I}\mathcal{A}_i}=\bigcup_{i\in I}\overline{\mathcal{A}}_i,$$

and if $\mathcal{B}$ is an event, then the Boolean distributive law generalizes as follows:

$$\mathcal{B}\cap\Big(\bigcup_{i\in I}\mathcal{A}_i\Big)=\bigcup_{i\in I}(\mathcal{B}\cap\mathcal{A}_i) \quad\text{and}\quad \mathcal{B}\cup\Big(\bigcap_{i\in I}\mathcal{A}_i\Big)=\bigcap_{i\in I}(\mathcal{B}\cup\mathcal{A}_i).$$

We now generalize (3) and (4) to families of events. Let $\{\mathcal{A}_i\}_{i\in I}$ be a finite family of events (i.e., the index set I is finite). Using (3), it follows by induction on $|I|$ that

$$\Pr\Big[\bigcup_{i\in I}\mathcal{A}_i\Big]\le\sum_{i\in I}\Pr[\mathcal{A}_i], \tag{6}$$

which is known as **Boole's inequality** (and sometimes called the **union bound**). Analogously, using (4), it follows by induction on $|I|$ that

$$\Pr\Big[\bigcup_{i\in I}\mathcal{A}_i\Big]=\sum_{i\in I}\Pr[\mathcal{A}_i] \quad\text{if } \{\mathcal{A}_i\}_{i\in I} \text{ is pairwise disjoint}, \tag{7}$$

that is, if $\mathcal{A}_i\cap\mathcal{A}_j=\emptyset$ for all $i,j\in I$ with $i\neq j$. We shall refer to (7) as **Boole's equality**. Both (6) and (7) are invaluable tools in calculating or estimating the probability of an event $\mathcal{A}$ by breaking $\mathcal{A}$ up into a family $\{\mathcal{A}_i\}_{i\in I}$ of smaller, and hopefully simpler, events, whose union is $\mathcal{A}$. We shall make frequent use of them.

We next consider a useful way to "glue together" probability distributions. Suppose one conducts two physically separate and unrelated random experiments, with each experiment modeled separately as a probability distribution. What we would like is a way to combine these distributions, obtaining a single probability distribution that models the two experiments as one grand experiment. This can be accomplished as follows.

Let $\Pr_1:\Omega_1\to[0,1]$ and $\Pr_2:\Omega_2\to[0,1]$ be probability distributions. Their **product distribution** $\Pr:=\Pr_1\Pr_2$ is defined as follows:

$$\begin{aligned}\Pr:\ &\Omega_1\times\Omega_2\to[0,1]\\ &(\omega_1,\omega_2)\mapsto\Pr_1(\omega_1)\Pr_2(\omega_2).\end{aligned}$$

It is easily verified that Pr is a probability distribution on the sample space $\Omega_1\times\Omega_2$:

$$\sum_{\omega_1,\omega_2}\Pr(\omega_1,\omega_2)=\sum_{\omega_1,\omega_2}\Pr_1(\omega_1)\Pr_2(\omega_2)=\Big(\sum_{\omega_1}\Pr_1(\omega_1)\Big)\Big(\sum_{\omega_2}\Pr_2(\omega_2)\Big)=1\cdot 1=1.$$

More generally, if $\Pr_i:\Omega_i\to[0,1]$, for $i=1,\ldots,n$, are probability distributions, then their product distribution is $\Pr:=\Pr_1\cdots\Pr_n$, where

$$\begin{aligned}\Pr:\ &\Omega_1\times\cdots\times\Omega_n\to[0,1]\\ &(\omega_1,\ldots,\omega_n)\mapsto\Pr_1(\omega_1)\cdots\Pr_n(\omega_n).\end{aligned}$$

If $\Pr_1=\Pr_2=\cdots=\Pr_n$, then we may write $\Pr=\Pr_1^n$. It is clear from the definitions that if each $\Pr_i$ is the uniform distribution on Ω_i, then Pr is the uniform distribution on $\Omega_1\times\cdots\times\Omega_n$.

Example 7. We can view the probability distribution Pr in Example 6 as $\Pr_1^2$, where $\Pr_1$ is the uniform distribution on $\{1, \ldots, 6\}$. $\square$

Example 8. Suppose we have a coin that comes up *heads* with some probability p, and *tails* with probability $q := 1 - p$. We toss the coin n times, and record the outcomes. We can model this as the product distribution $\Pr = \Pr_1^n$, where $\Pr_1$ is the distribution of a Bernoulli trial (see Example 3) with success probability p, and where we identify *success* with *heads*, and *failure* with *tails*. The sample space Ω of Pr is the set of all 2^n tuples $\omega = (\omega_1, \ldots, \omega_n)$, where each ω_i is either *heads* or *tails*. If the tuple ω has k *heads* and $n - k$ *tails*, then $\Pr(\omega) = p^k q^{n-k}$, regardless of the positions of the *heads* and *tails* in the tuple.

For each $k = 0, \ldots, n$, let $\mathcal{A}_k$ be the event that our coin comes up *heads* exactly k times. As a set, $\mathcal{A}_k$ consists of all those tuples in the sample space with exactly k *heads*, and so

$$|\mathcal{A}_k| = \binom{n}{k},$$

from which it follows that

$$\Pr[\mathcal{A}_k] = \binom{n}{k} p^k q^{n-k}.$$

If our coin is a fair coin, so that $p = q = 1/2$, then Pr is the uniform distribution on Ω, and for each $k = 0, \ldots, n$, we have

$$\Pr[\mathcal{A}_k] = \binom{n}{k} 2^{-n}. \quad \square$$

The previous example made use of binomial coefficients, which the reader may wish to review in §A2.

Suppose $\Pr : \Omega \to [0, 1]$ is a probability distribution. Now consider another probability distribution $\Pr' : \Omega' \to [0, 1]$. Of course, these two distributions are equal if and only if $\Omega = \Omega'$ and $\Pr(\omega) = \Pr'(\omega)$ for all $\omega \in \Omega$. However, it is natural and convenient to have a more relaxed notion of equality. We shall say that Pr and $\Pr'$ are **essentially equal** if they are equal on all outcomes in $\Omega \cap \Omega'$ and zero everywhere else. For example, if Pr is the probability distribution on $\{1, 2, 3, 4\}$ that assigns probability 1/3 to 1, 2, and 3, and probability 0 to 4, we may say that Pr is essentially the uniform distribution on $\{1, 2, 3\}$.

2 Conditional probability and independence

Let Pr be a probability distribution on a sample space Ω.

For a given event $\mathcal{B} \subseteq \Omega$ with $\Pr[\mathcal{B}] \neq 0$, and for $\omega \in \Omega$, let us define

$$\Pr(\omega \mid \mathcal{B}) := \begin{cases} \Pr(\omega)/\Pr[\mathcal{B}] & \text{if } \omega \in \mathcal{B}, \\ 0 & \text{otherwise.} \end{cases}$$

Viewing $\mathcal{B}$ as fixed, the function $\Pr(\cdot \mid \mathcal{B})$ is a new probability distribution on the sample space Ω, called the **conditional distribution (derived from** Pr**) given** $\mathcal{B}$.

Intuitively, $\Pr(\cdot \mid \mathcal{B})$ has the following interpretation. Suppose a random experiment produces an outcome according to the distribution Pr. Further, suppose we learn that the event $\mathcal{B}$ has occurred, but nothing else about the outcome. Then the distribution $\Pr(\cdot \mid \mathcal{B})$ assigns new probabilities to all possible outcomes, reflecting the partial knowledge that the event $\mathcal{B}$ has occurred.

For a given event $\mathcal{A} \subseteq \Omega$, its probability with respect to the conditional distribution given $\mathcal{B}$ is

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \sum_{\omega \in \mathcal{A}} \Pr(\omega \mid \mathcal{B}) = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{B}]}.$$

The value $\Pr[\mathcal{A} \mid \mathcal{B}]$ is called the **conditional probability of $\mathcal{A}$ given $\mathcal{B}$**. Again, the intuition is that this is the probability that the event $\mathcal{A}$ occurs, given the partial knowledge that the event $\mathcal{B}$ has occurred.

For events $\mathcal{A}$ and $\mathcal{B}$, if $\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \Pr[\mathcal{B}]$, then $\mathcal{A}$ and $\mathcal{B}$ are called **independent** events. If $\Pr[\mathcal{B}] \neq 0$, one easily sees that $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if $\Pr[\mathcal{A} \mid \mathcal{B}] = \Pr[\mathcal{A}]$; intuitively, independence means that the partial knowledge that event $\mathcal{B}$ has occurred does not affect the likelihood that $\mathcal{A}$ occurs.

Example 9. Suppose Pr is the uniform distribution on Ω, and that $\mathcal{B} \subseteq \Omega$ with $\Pr[\mathcal{B}] \neq 0$. Then the conditional distribution given $\mathcal{B}$ is essentially the uniform distribution on $\mathcal{B}$. □

Example 10. Consider again Example 4, where $\mathcal{A}$ is the event that the value on the die is odd, and $\mathcal{B}$ is the event that the value of the die exceeds 2. Then as we calculated, $\Pr[\mathcal{A}] = 1/2$, $\Pr[\mathcal{B}] = 2/3$, and $\Pr[\mathcal{A} \cap \mathcal{B}] = 1/3$; thus, $\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \Pr[\mathcal{B}]$, and we conclude that $\mathcal{A}$ and $\mathcal{B}$ are independent. Indeed, $\Pr[\mathcal{A} \mid \mathcal{B}] = (1/3)/(2/3) = 1/2 = \Pr[\mathcal{A}]$; intuitively, given the partial knowledge that the value on the die exceeds 2, we know it is equally likely to be either 3, 4, 5, or 6, and so the conditional probability that it is odd is 1/2.

However, consider the event $\mathcal{C}$ that the value on the die exceeds 3. We have $\Pr[\mathcal{C}] = 1/2$ and $\Pr[\mathcal{A} \cap \mathcal{C}] = 1/6 \neq 1/4$, from which we conclude that $\mathcal{A}$ and $\mathcal{C}$ are *not* independent. Indeed, $\Pr[\mathcal{A} \mid \mathcal{C}] = (1/6)/(1/2) = 1/3 \neq \Pr[\mathcal{A}]$; intuitively, given the partial knowledge that the value on the die exceeds 3, we know it is equally likely to be either 4, 5, or 6, and so the conditional probability that it is odd is just 1/3, and not 1/2. □

Example 11. In Example 6, suppose that Alice tells Bob the sum of the two dice before Bob makes his guess. The following table is useful for visualizing the situation:

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	2	3	4	5	6

For example, suppose Alice tells Bob the sum is 4. Then what is Bob's best strategy in this case? Let $\mathcal{D}_\ell$ be the event that the sum is ℓ, for $\ell = 2, \ldots, 12$, and consider the conditional distribution given $\mathcal{D}_4$. This conditional distribution is essentially the uniform distribution on the set $\{(1,3),(2,2),(3,1)\}$. The numbers 1 and 3 both appear in two pairs, while the number 2 appears in just one pair. Therefore,

$$\Pr[\mathcal{C}_1 \mid \mathcal{D}_4] = \Pr[\mathcal{C}_3 \mid \mathcal{D}_4] = 2/3,$$

while

$$\Pr[\mathcal{C}_2 \mid \mathcal{D}_4] = 1/3$$

and

$$\Pr[\mathcal{C}_4 \mid \mathcal{D}_4] = \Pr[\mathcal{C}_5 \mid \mathcal{D}_4] = \Pr[\mathcal{C}_6 \mid \mathcal{D}_4] = 0.$$

Thus, if the sum is 4, Bob's best strategy is to guess either 1 or 3, which will be correct with probability 2/3. Clearly, for each $k = 1, \ldots, 6$, the events $\mathcal{C}_k$ and $\mathcal{D}_4$ are *not* independent.

Similarly, if the sum is 5, then we consider the conditional distribution given $\mathcal{D}_5$, which is essentially the uniform distribution on $\{(1,4),(2,3),(3,2),(4,1)\}$. In this case, Bob should choose one of the numbers $k = 1, \ldots, 4$, each of which will be correct with probability $\Pr[\mathcal{C}_k \mid \mathcal{D}_5] = 1/2$. □

Suppose $\{\mathcal{B}_i\}_{i \in I}$ is a finite, pairwise disjoint family of events, whose union is Ω. Now consider an arbitrary event $\mathcal{A}$. Since $\{\mathcal{A} \cap \mathcal{B}_i\}_{i \in I}$ is a pairwise disjoint family of events whose union is $\mathcal{A}$, Boole's equality (7) implies

$$\Pr[\mathcal{A}] = \sum_{i \in I} \Pr[\mathcal{A} \cap \mathcal{B}_i]. \tag{8}$$

Furthermore, we have

$$\Pr[\mathcal{A}] = \sum_{i \in I} \Pr[\mathcal{A} \mid \mathcal{B}_i] \Pr[\mathcal{B}_i], \tag{9}$$

with the understanding that if any of the $\mathcal{B}_i$'s occur with zero probability, the corresponding terms in (9) are excluded. Continuing, if $\Pr[\mathcal{A}] \neq 0$, then for each $j \in I$ with $\Pr[\mathcal{B}_j] \neq 0$, we have

$$\Pr[\mathcal{B}_j \mid \mathcal{A}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}_j]}{\Pr[\mathcal{A}]} = \frac{\Pr[\mathcal{A} \mid \mathcal{B}_j] \Pr[\mathcal{B}_j]}{\sum_{i \in I} \Pr[\mathcal{A} \mid \mathcal{B}_i] \Pr[\mathcal{B}_i]}, \tag{10}$$

where, again, we exclude any terms in the sum for which $\Pr[\mathcal{B}_i]$ is zero. Equations (8) and (9) are sometimes called the **law of total probability**, while equation (10) is known as **Bayes' theorem**. Equation (9) (resp., (10)) is useful for computing or estimating $\Pr[\mathcal{A}]$ (resp., $\Pr[\mathcal{B}_j \mid \mathcal{A}]$) by conditioning on the events $\mathcal{B}_i$.

Example 12. Let us continue with Example 11, and compute Bob's overall probability of winning, assuming he follows an optimal strategy. If the sum is 2 or 12, clearly there is only one sensible choice for Bob to make, and it will certainly be correct. If the sum is any other number ℓ, and there are N_ℓ pairs in the sample space that sum to that number, then there will always be a value that appears in exactly 2 of these N_ℓ pairs, and Bob should choose such a value (see the diagram in Example 11). Indeed, this is achieved by the simple rule of choosing the value 1 if $\ell \leq 7$, and the value 6 if $\ell > 7$. This is an optimal strategy for Bob, and if $\mathcal{C}$ is the event that Bob wins following this strategy, then by total probability (9), we have

$$\Pr[\mathcal{C}] = \sum_{\ell=2}^{12} \Pr[\mathcal{C} \mid \mathcal{D}_\ell] \Pr[\mathcal{D}_\ell].$$

Moreover,

$$\Pr[\mathcal{C} \mid \mathcal{D}_2] \Pr[\mathcal{D}_2] = 1 \cdot \frac{1}{36} = \frac{1}{36}, \quad \Pr[\mathcal{C} \mid \mathcal{D}_{12}] \Pr[\mathcal{D}_{12}] = 1 \cdot \frac{1}{36} = \frac{1}{36},$$

and for $\ell = 3, \ldots, 11$, we have

$$\Pr[\mathcal{C} \mid \mathcal{D}_\ell] \Pr[\mathcal{D}_\ell] = \frac{2}{N_\ell} \cdot \frac{N_\ell}{36} = \frac{1}{18}.$$

Therefore,

$$\Pr[\mathcal{C}] = \frac{1}{36} + \frac{1}{36} + \frac{9}{18} = \frac{10}{18}. \quad \square$$

Example 13. Suppose that the rate of incidence of disease X in the overall population is 1%. Also suppose that there is a test for disease X; however, the test is not perfect: while all sick patients test positive for the disease, a healthy patient may test positive with a 5% probability (we say the test has a 5% false positive rate and a 0% false negative rate, in this case). A doctor gives the test to a patient and it comes out positive. How should the doctor advise his patient? In particular, what is the probability that the patient actually has disease X, given a positive test result?

Amazingly, many trained doctors will say the probability is 95%, since the test has a false positive rate of 5%. However, this conclusion is completely wrong.

Let $\mathcal{A}$ be the event that the test is positive and let $\mathcal{B}$ be the event that the patient has disease X. The relevant quantity that we need to estimate is $\Pr[\mathcal{B} \mid \mathcal{A}]$; that is, the probability that the patient has disease X, given a positive test result. We use Bayes' theorem to do this:

$$\Pr[\mathcal{B} \mid \mathcal{A}] = \frac{\Pr[\mathcal{A} \mid \mathcal{B}] \Pr[\mathcal{B}]}{\Pr[\mathcal{A} \mid \mathcal{B}] \Pr[\mathcal{B}] + \Pr[\mathcal{A} \mid \overline{\mathcal{B}}] \Pr[\overline{\mathcal{B}}]} = \frac{1 \cdot 0.01}{1 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.17.$$

Thus, the chances that the patient has disease X given a positive test result are just 17%. The correct intuition here is that it is much more likely to get a false positive than it is to actually have the disease.

Of course, the real world is a bit more complicated than this example suggests: the doctor may be giving the patient the test because other risk factors or symptoms may suggest that the patient is more likely to have the disease than a random member of the population, in which case the above analysis does not apply. □

Example 14. This example is based on the TV game show "Let's make a deal," which was popular in the 1960's and 1970's, and has been resurrected again in recent years. In this game, a contestant chooses one of three doors. Behind two doors is a "zonk," that is, something amusing but of little or no value, such as a goat, and behind one of the doors is a "grand prize," such as a car or vacation package. We may assume that the door behind which the grand prize is placed is chosen at random from among the three doors, with equal probability. After the contestant chooses a door, the host of the show always reveals a zonk behind one of the two doors not chosen by the contestant. The contestant is then given a choice: either stay with his initial choice of door, or switch to the other unopened door. After the contestant finalizes his decision on which door to choose, that door is opened and he wins whatever is behind it. The question is, which strategy is better for the contestant: to stay or to switch?

Let us evaluate the two strategies. If the contestant always stays with his initial selection, then it is clear that his probability of success is exactly 1/3.

Now consider the strategy of always switching. Let $\mathcal{B}$ be the event that the contestant's initial choice was correct, and let $\mathcal{A}$ be the event that the contestant wins the grand prize. On the one hand, if the contestant's initial choice was correct, then switching will certainly lead to failure (in this case, the host has two doors to choose from, but his choice does not affect the outcome). Thus, $\Pr[\mathcal{A} \mid \mathcal{B}] = 0$. On the other hand, suppose that the contestant's initial choice was incorrect, so that one of the zonks is behind the initially chosen door. Since the host reveals the other zonk, switching will lead with certainty to success. Thus, $\Pr[\mathcal{A} \mid \overline{\mathcal{B}}] = 1$. Furthermore, it is clear that $\Pr[\mathcal{B}] = 1/3$. So using total probability (9), we compute

$$\Pr[\mathcal{A}] = \Pr[\mathcal{A} \mid \mathcal{B}] \Pr[\mathcal{B}] + \Pr[\mathcal{A} \mid \overline{\mathcal{B}}] \Pr[\overline{\mathcal{B}}] = 0 \cdot (1/3) + 1 \cdot (2/3) = 2/3.$$

Thus, the "stay" strategy has a success probability of 1/3, while the "switch" strategy has a success probability of 2/3. So it is better to switch than to stay.

Of course, real life is a bit more complicated. The host does not always reveal a zonk and offer a choice to switch. Indeed, if the host *only* revealed a zonk when the contestant had chosen the correct door, then switching would certainly be the wrong strategy. However, if the host's choice itself were a random decision made independently of the contestant's initial choice, then switching is again the preferred strategy. □

We close this section with a simple fact about independent events and their complements.

Theorem 1. *If $\mathcal{A}$ and $\mathcal{B}$ are independent events, then so are $\mathcal{A}$ and $\overline{\mathcal{B}}$.*

Proof. We have

$$\begin{aligned}\Pr[\mathcal{A}] &= \Pr[\mathcal{A} \cap \mathcal{B}] + \Pr[\mathcal{A} \cap \overline{\mathcal{B}}] \quad \text{(by total probability (8))} \\ &= \Pr[\mathcal{A}]\Pr[\mathcal{B}] + \Pr[\mathcal{A} \cap \overline{\mathcal{B}}] \quad \text{(since } \mathcal{A} \text{ and } \mathcal{B} \text{ are independent).}\end{aligned}$$

Therefore,

$$\Pr[\mathcal{A} \cap \overline{\mathcal{B}}] = \Pr[\mathcal{A}] - \Pr[\mathcal{A}]\Pr[\mathcal{B}] = \Pr[\mathcal{A}](1 - \Pr[\mathcal{B}]) = \Pr[\mathcal{A}]\Pr[\overline{\mathcal{B}}]. \quad \square$$

This theorem implies that

$$\begin{aligned}\mathcal{A} \text{ and } \mathcal{B} \text{ are independent} &\iff \mathcal{A} \text{ and } \overline{\mathcal{B}} \text{ are independent} \\ &\iff \overline{\mathcal{A}} \text{ and } \mathcal{B} \text{ " " } \\ &\iff \overline{\mathcal{A}} \text{ and } \overline{\mathcal{B}} \text{ " " } .\end{aligned}$$

3 Random variables

It is sometimes convenient to associate a real number, or other mathematical object, with each outcome of a random experiment. The notion of a random variable formalizes this idea.

Let Pr be a probability distribution on a sample space Ω. A **random variable** X is a function $\mathsf{X} : \Omega \to S$, where S is some set, and we say that X **takes values in** S. We do not require that the values taken by X are real numbers, but if this is the case, we say that X is **real valued**. For $s \in S$, "$\mathsf{X} = s$" denotes the event $\{\omega \in \Omega : \mathsf{X}(\omega) = s\}$. It is immediate from this definition that

$$\Pr[\mathsf{X} = s] = \sum_{\omega \in \mathsf{X}^{-1}(\{s\})} \Pr(\omega).$$

More generally, for any predicate ϕ on S, we may write "$\phi(\mathsf{X})$" as shorthand for the event $\{\omega \in \Omega : \phi(\mathsf{X}(\omega))\}$. When we speak of the **image** of X, we simply mean its image in the usual function-theoretic sense, that is, the set $\mathsf{X}(\Omega) = \{\mathsf{X}(\omega) : \omega \in \Omega\}$. While a random variable is simply a function on the sample space, any discussion of its properties always takes place relative to a particular probability distribution, which may be implicit from context.

One can easily combine random variables to define new random variables. Suppose $\mathsf{X}_1, \ldots, \mathsf{X}_n$ are random variables, where $\mathsf{X}_i : \Omega \to S_i$ for $i = 1, \ldots, n$. Then $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ denotes the random variable that maps $\omega \in \Omega$ to $(\mathsf{X}_1(\omega), \ldots, \mathsf{X}_n(\omega)) \in S_1 \times \cdots \times S_n$. If $f : S_1 \times \cdots \times S_n \to T$ is a function, then $f(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ denotes the random variable that maps $\omega \in \Omega$ to $f(\mathsf{X}_1(\omega), \ldots, \mathsf{X}_n(\omega))$. If f is applied using a special notation, the same notation may be applied to denote the resulting random variable; for example, if X and Y are random variables taking values in a set S, and $\star$ is a binary operation on S, then $\mathsf{X} \star \mathsf{Y}$ denotes the random variable that maps $\omega \in \Omega$ to $\mathsf{X}(\omega) \star \mathsf{Y}(\omega) \in S$.

Let X be a random variable whose image is S. The variable X determines a probability distribution $\Pr_{\mathsf{X}} : S \to [0,1]$ on the set S, where $\Pr_{\mathsf{X}}(s) := \Pr[\mathsf{X} = s]$ for each $s \in S$. We call $\Pr_{\mathsf{X}}$ the **distribution of** X. If $\Pr_{\mathsf{X}}$ is the uniform distribution on S, then we say that X is **uniformly distributed over** S.

Suppose X and Y are random variables that take values in a set S. If $\Pr[\mathsf{X} = s] = \Pr[\mathsf{Y} = s]$ for all $s \in S$, then the distributions of X and Y are essentially equal even if their images are not identical.

Example 15. Again suppose we roll two dice, and model this experiment as the uniform distribution on $\Omega := \{1,\ldots,6\} \times \{1,\ldots,6\}$. We can define the random variable X that takes the value of the first die, and the random variable Y that takes the value of the second; formally, X and Y are functions on Ω, where

$$\mathsf{X}(s,t) := s \quad \text{and} \quad \mathsf{Y}(s,t) := t \quad \text{for } (s,t) \in \Omega.$$

For each value $s \in \{1,\ldots,6\}$, the event $\mathsf{X} = s$ is $\{(s,1),\ldots,(s,6)\}$, and so $\Pr[\mathsf{X} = s] = 6/36 = 1/6$. Thus, X is uniformly distributed over $\{1,\ldots,6\}$. Likewise, Y is uniformly distributed over $\{1,\ldots,6\}$, and the random variable (X,Y) is uniformly distributed over Ω. We can also define the random variable $\mathsf{Z} := \mathsf{X} + \mathsf{Y}$, which formally is the function on the sample space defined by

$$\mathsf{Z}(s,t) := s + t \quad \text{for } (s,t) \in \Omega.$$

The image of Z is $\{2,\ldots,12\}$, and its distribution is given by the following table:

u	2	3	4	5	6	7	8	9	10	11	12
$\Pr[\mathsf{Z} = u]$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

. □

Example 16. If $\mathcal{A}$ is an event, we may define a random variable X as follows: $\mathsf{X} := 1$ if the event $\mathcal{A}$ occurs, and $\mathsf{X} := 0$ otherwise. The variable X is called the **indicator variable for** $\mathcal{A}$. Formally, X is the function that maps $\omega \in \mathcal{A}$ to 1, and $\omega \in \Omega \setminus \mathcal{A}$ to 0; that is, X is simply the characteristic function of $\mathcal{A}$. The distribution of X is that of a Bernoulli trial: $\Pr[\mathsf{X} = 1] = \Pr[\mathcal{A}]$ and $\Pr[\mathsf{X} = 0] = 1 - \Pr[\mathcal{A}]$.

It is not hard to see that $1 - \mathsf{X}$ is the indicator variable for $\overline{\mathcal{A}}$. Now suppose $\mathcal{B}$ is another event, with indicator variable Y. Then it is also not hard to see that XY is the indicator variable for $\mathcal{A} \cap \mathcal{B}$, and that $\mathsf{X} + \mathsf{Y} - \mathsf{XY}$ is the indicator variable for $\mathcal{A} \cup \mathcal{B}$; in particular, if $\mathcal{A} \cap \mathcal{B} = \emptyset$, then $\mathsf{X} + \mathsf{Y}$ is the indicator variable for $\mathcal{A} \cup \mathcal{B}$. □

Example 17. Consider again Example 8, where we have a coin that comes up *heads* with probability p, and *tails* with probability $q := 1 - p$, and we toss it n times. For each $i = 1,\ldots,n$, let $\mathcal{A}_i$ be the event that the ith toss comes up *heads*, and let X_i be the corresponding indicator variable. Let us also define $\mathsf{X} := \mathsf{X}_1 + \cdots + \mathsf{X}_n$, which represents the total number of tosses that come up *heads*. The image of X is $\{0,\ldots,n\}$. By the calculations made in Example 8, for each $k = 0,\ldots,n$, we have

$$\Pr[\mathsf{X} = k] = \binom{n}{k} p^k q^{n-k}.$$

The distribution of the random variable X is called a **binomial distribution**. Such a distribution is parameterized by the success probability p of the underlying Bernoulli trial, and by the number of times n the trial is repeated. □

Uniform distributions are very nice, simple distributions. It is therefore good to have simple criteria that ensure that certain random variables have uniform distributions. The next theorem provides one such criterion. We need a definition: if S and T are finite sets, then we say that a

given function $f : S \to T$ is a **regular function** if every element in the image of f has the same number of pre-images under f. An injective function is a special case of a regular function.

Theorem 2. *Suppose $f : S \to T$ is a surjective, regular function, and that X is a random variable that is uniformly distributed over S. Then $f(\mathsf{X})$ is uniformly distributed over T.*

Proof. The assumption that f is surjective and regular implies that for every $t \in T$, the set $S_t := f^{-1}(\{t\})$ has size $|S|/|T|$. So, for each $t \in T$, working directly from the definitions, we have

$$\begin{aligned}\Pr[f(\mathsf{X}) = t] &= \sum_{\omega \in \mathsf{X}^{-1}(S_t)} \Pr(\omega) = \sum_{s \in S_t} \sum_{\omega \in \mathsf{X}^{-1}(\{s\})} \Pr(\omega) = \sum_{s \in S_t} \Pr[\mathsf{X} = s] \\ &= \sum_{s \in S_t} 1/|S| = (|S|/|T|)/|S| = 1/|T|. \quad \square\end{aligned}$$

Let X be a random variable whose image is S. Let $\mathcal{B}$ be an event with $\Pr[\mathcal{B}] \neq 0$. The **conditional distribution of X given $\mathcal{B}$** is defined to be the distribution of X *relative to the conditional distribution* $\Pr(\cdot \mid \mathcal{B})$, that is, the distribution $\Pr_{\mathsf{X}|\mathcal{B}} : S \to [0, 1]$ defined by $\Pr_{\mathsf{X}|\mathcal{B}}(s) := \Pr[\mathsf{X} = s \mid \mathcal{B}]$ for $s \in S$.

Suppose X and Y are random variables, with images S and T, respectively. We say X and Y are **independent** if for all $s \in S$ and all $t \in T$, the events $\mathsf{X} = s$ and $\mathsf{Y} = t$ are independent, which is to say,

$$\Pr[(\mathsf{X} = s) \cap (\mathsf{Y} = t)] = \Pr[\mathsf{X} = s] \Pr[\mathsf{Y} = t].$$

Equivalently, X and Y are independent if and only if the distribution of (X, Y) is essentially equal to the product of the distribution of X and the distribution of Y. As a special case, if X is uniformly distributed over S, and Y is uniformly distributed over T, then X and Y are independent if and only if (X, Y) is uniformly distributed over $S \times T$.

Independence can also be characterized in terms of conditional probabilities. From the definitions, it is immediate that X and Y are independent if and only if for all values t taken by Y with non-zero probability, we have

$$\Pr[\mathsf{X} = s \mid \mathsf{Y} = t] = \Pr[\mathsf{X} = s]$$

for all $s \in S$; that is, the conditional distribution of X given $\mathsf{Y} = t$ is the same as the distribution of X. From this point of view, an intuitive interpretation of independence is that information about the value of one random variable does not reveal any information about the value of the other.

Example 18. Let us continue with Example 15. The random variables X and Y are independent: each is uniformly distributed over $\{1, \ldots, 6\}$, and (X, Y) is uniformly distributed over $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$. Let us calculate the conditional distribution of X given $\mathsf{Z} = 4$. We have $\Pr[\mathsf{X} = s \mid \mathsf{Z} = 4] = 1/3$ for $s = 1, 2, 3$, and $\Pr[\mathsf{X} = s \mid \mathsf{Z} = 4] = 0$ for $s = 4, 5, 6$. Thus, the conditional distribution of X given $\mathsf{Z} = 4$ is essentially the uniform distribution on $\{1, 2, 3\}$. Let us calculate the conditional distribution of Z given $\mathsf{X} = 1$. We have $\Pr[\mathsf{Z} = u \mid \mathsf{X} = 1] = 1/6$ for $u = 2, \ldots, 7$, and $\Pr[\mathsf{Z} = u \mid \mathsf{X} = 1] = 0$ for $u = 8, \ldots, 12$. Thus, the conditional distribution of Z given $\mathsf{X} = 1$ is essentially the uniform distribution on $\{2, \ldots, 7\}$. In particular, it is clear that X and Z are *not* independent. $\square$

Example 19. Let m be a positive integer. Suppose X and Y are independent random variables, with each uniformly distributed over $\mathbb{Z}_m$ (the integers mod m). This means that (X, Y) is uniformly distributed over $\mathbb{Z}_m \times \mathbb{Z}_m$. Consider $\mathsf{Z} := \mathsf{X} + \mathsf{Y}$ (the addition is in $\mathbb{Z}_m$, so $\mathsf{Z} \in \mathbb{Z}_m$).

First, we claim that Z is uniformly distributed over $\mathbb{Z}_m$. We can show this using Theorem 2 by showing that the function $f : \mathbb{Z}_m \times \mathbb{Z}_m \to \mathbb{Z}_m$ defined by $f(s, t) := s + t$ is surjective and regular.

Let $\alpha \in \mathbb{Z}_m$ be fixed. It will suffice to show that the number of solutions $(s,t) \in \mathbb{Z}_m \times \mathbb{Z}_m$ to the equation

$$s + t = \alpha \tag{11}$$

is equal to m. But this is clearly the case: for each choice of $s \in \mathbb{Z}_m$, there is a unique $t \in \mathbb{Z}_m$ satisfying (11), namely, $t := \alpha - s$.

Next, we claim that X and Z are independent. Let $\alpha, \beta \in \mathbb{Z}_m$ be fixed. We want to show that $\Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Z} = \beta)] = 1/m^2$. This can be seen as follows:

$$\begin{aligned}\Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Z} = \beta)] &= \Pr[(\mathsf{X} = \alpha) \cap (\mathsf{X} + \mathsf{Y} = \beta)] = \Pr[(\mathsf{X} = \alpha) \cap (\alpha + \mathsf{Y} = \beta)] \\ &= \Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Y} = \beta - \alpha)] \\ &= \Pr[\mathsf{X} = \alpha] \cdot \Pr[\mathsf{Y} = \beta - \alpha] \quad \text{(independence of } \mathsf{X} \text{ and } \mathsf{Y}) \\ &= \frac{1}{m} \cdot \frac{1}{m} = \frac{1}{m^2}. \ \square\end{aligned}$$

Example 20. As in the previous example, let m be a positive integer, and suppose X and Y are independent random variables taking values in $\mathbb{Z}_m$. We shall assume that Y is uniformly distributed over $\mathbb{Z}_m$. However, unlike in the previous example, we shall not assume anything about the distribution of X. We can nevertheless show that $\mathsf{Z} := \mathsf{X} + \mathsf{Y}$ are independent.

This has applications to cryptography. Suppose Y represents an encryption key shared between Alice and Bob. Alice encrypts a message X by computing the ciphertext $\mathsf{Z} = \mathsf{X} + \mathsf{Y}$ and sends Z over an insecure network. Bob can decrypt the ciphertext by computing $\mathsf{X} = \mathsf{Z} - \mathsf{Y}$. However, the independence of Z and X ensures that an eavesdropper who only learns the value of the ciphertext Z learns nothing about the message X.

We now prove independence. Let $\alpha, \beta \in \mathbb{Z}_m$ be fixed. As in the previous example, we have

$$\begin{aligned}\Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Z} = \beta)] &= \Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Y} = \beta - \alpha)] \\ &= \Pr[\mathsf{X} = \alpha] \cdot \Pr[\mathsf{Y} = \beta - \alpha] \quad \text{(independence of } \mathsf{X} \text{ and } \mathsf{Y}) \\ &= \Pr[\mathsf{X} = \alpha] \cdot \frac{1}{m}.\end{aligned}$$

So to prove independence, it will suffice to show that Z is uniformly distributed over $\mathbb{Z}_m$. We can show this using total probability and the equality we just proved. Let $\beta \in \mathbb{Z}_m$ be fixed. We have

$$\begin{aligned}\Pr[\mathsf{Z} = \beta] &= \sum_{\alpha \in \mathbb{Z}_m} \Pr[(\mathsf{X} = \alpha) \cap (\mathsf{Z} = \beta)] = \sum_{\alpha \in \mathbb{Z}_m} \Pr[\mathsf{X} = \alpha] \cdot \frac{1}{m} \\ &= \frac{1}{m} \sum_{\alpha \in \mathbb{Z}_m} \Pr[\mathsf{X} = \alpha] \\ &= \frac{1}{m} \quad \text{(probabilities sum to 1).} \ \square\end{aligned}$$

Example 21. Suppose $\mathcal{A}$ and $\mathcal{B}$ are events with corresponding indicator variables X and Y. Theorem 1 implies that $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if X and Y are independent. $\square$

We now generalize the notion of independence to families of random variables. Intuitively, the notion of independence we define means that revealing information about the values of some variables in the family does not reveal any information about other variables in the family.

Let $\{\mathsf{X}_i\}_{i \in I}$ be a finite family of random variables. Let us call a corresponding family of values $\{s_i\}_{i \in I}$ an **assignment** to $\{\mathsf{X}_i\}_{i \in I}$ if s_i is in the image of X_i for each $i \in I$. The family $\{\mathsf{X}_i\}_{i \in I}$ is

called **mutually independent** if for every assignment $\{s_i\}_{i\in I}$ to $\{\mathsf{X}_i\}_{i\in I}$, we have

$$\Pr\Big[\bigcap_{i\in I}(\mathsf{X}_i = s_i)\Big] = \prod_{i\in I}\Pr[\mathsf{X}_i = s_j].$$

Also, for $k \le |I|$, we say that $\{\mathsf{X}_i\}_{i\in I}$ is **k-wise independent** if $\{\mathsf{X}_j\}_{j\in J}$ is mutually independent for every subset $J \subseteq I$ of size k. We say $\{\mathsf{X}_i\}_{i\in I}$ is **pairwise independent** if it is 2-wise independent.

Example 22. As in Example 19, let X and Y be independent random variables, with each uniformly distributed over $\mathbb{Z}_m$, and define $\mathsf{Z} := \mathsf{X} + \mathsf{Y}$. In that example, we showed that Z is also uniformly distributed over $\mathbb{Z}_m$. We also showed that X and Z are independent. The same argument shows that Y and Z are independent. It follows that the family of random variables X, Y, Z is pairwise independent. However, it is not mutually independent. For example,

$$\Pr[(\mathsf{X} = 0) \cap (\mathsf{Y} = 0) \cap (\mathsf{Z} = 1)] = 0 \neq 1/m^3. \quad \square$$

We next state and prove a simple theorem that says k-wise independence implies ℓ-wise independence for any $\ell < k$.

Theorem 3. *Suppose $\{\mathsf{X}_i\}_{i\in I}$ is k-wise independent, where $1 < k \le |I|$. Then it is also ℓ-wise independent for any ℓ with $1 < \ell < k$.*

Proof. It suffices to prove the theorem for $\ell := k - 1$, as repeated application gives the result for any ℓ.

We want to show that

$$\Pr[(\mathsf{X}_{i_1} = s_1) \cap \cdots \cap (\mathsf{X}_{i_\ell} = s_\ell)] = \Pr[\mathsf{X}_{i_1} = s_1] \cdots \Pr[\mathsf{X}_{i_\ell} = s_\ell]$$

for any set of ℓ distinct indices $i_1, \ldots, i_\ell$ and any assignment $s_1, \ldots, s_\ell$ to the variables $\mathsf{X}_{i_1}, \ldots, \mathsf{X}_{i_\ell}$. Let i_k be any index not among $i_1, \ldots, i_\ell$, and consider the random variable X_{i_k}. Using the law of total probability, if we sum over all s_{i_k} in the range of X_{i_k}, we have

$$\begin{aligned}
&\Pr[(\mathsf{X}_{i_1} = s_1) \cap \cdots \cap (\mathsf{X}_{i_\ell} = s_\ell)] \\
&\quad = \sum_{s_{i_k}} \Pr[(\mathsf{X}_{i_1} = s_1) \cap \cdots \cap (\mathsf{X}_{i_\ell} = s_\ell) \cap (\mathsf{X}_{i_k} = s_{i_k})] \\
&\quad = \sum_{s_{i_k}} \Pr[\mathsf{X}_{i_1} = s_1] \cdots \Pr[\mathsf{X}_{i_\ell} = s_\ell] \Pr[\mathsf{X}_{i_k} = s_{i_k}] \quad (k\text{-wise independence}) \\
&\quad = \Pr[\mathsf{X}_{i_1} = s_1] \cdots \Pr[\mathsf{X}_{i_\ell} = s_\ell] \sum_{s_{i_k}} \Pr[\mathsf{X}_{i_k} = s_{i_k}] \\
&\quad = \Pr[\mathsf{X}_{i_1} = s_1] \cdots \Pr[\mathsf{X}_{i_\ell} = s_\ell] \quad (\text{probabilities sum to 1}). \quad \square
\end{aligned}$$

From this theorem, we see that $\{\mathsf{X}_i\}_{i\in I}$ is mutually independent if and only it is k-wise independent for all $k = 2, \ldots, |I|$.

It is also useful to understand mutual independence in terms of conditional probabilities. We leave the proof of the following theorem to the reader.

Theorem 4. *The family $\{\mathsf{X}_i\}_{i=1}^n$ of random variables is mutually independent if and only if the following holds: for all $s_1, \ldots, s_n$, and for $i = 2, \ldots, n$, we have*

$$\Pr[\mathsf{X}_i = s_i \mid (\mathsf{X}_1 = s_1) \cap \cdots \cap (\mathsf{X}_{i-1} = s_{i-1})] = \Pr[\mathsf{X}_i = s_i],$$

provided $\Pr[(\mathsf{X}_1 = s_1) \cap \cdots \cap (\mathsf{X}_{i-1} = s_{i-1})] \neq 0$.

Intuitively, this says revealing the values of some random variables does not leak any information about any of the other random variables.

We next develop several results that can be used to establish independence.

An immediate consequence of Theorem 3, we obtain the following:

Theorem 5. *Suppose $\{\mathsf{X}_i\}_{i=1}^n$ is a family of random variables, and that m is an integer with $0 < m < n$. Then the following are equivalent:*

(i) $\{\mathsf{X}_i\}_{i=1}^n$ *is mutually independent;*

(ii) $\{\mathsf{X}_i\}_{i=1}^m$ *is mutually independent,* $\{\mathsf{X}_i\}_{i=m+1}^n$ *is mutually independent, and the two variables* $(\mathsf{X}_1, \ldots, \mathsf{X}_m)$ *and* $(\mathsf{X}_{m+1}, \ldots, \mathsf{X}_n)$ *are independent.*

The argument made in the proof of Theorem 3 can be adapted to obtain a more general condition which can be used to establish independence.

Theorem 6. *Let $\{\mathsf{X}_i\}_{i=1}^n$ be a family of random variables, where each X_i has image S_i. Also, let $\{f_i\}_{i=1}^n$ be a family of functions, where $f_i : S_i \to [0,1]$ and $\sum_{s_i \in S_i} f_i(s_i) = 1$ for each $i = 1, \ldots, n$. Further, suppose that*

$$\Pr[(\mathsf{X}_1 = s_1) \cap \cdots \cap (\mathsf{X}_n = s_n)] = f_1(s_1) \cdots f_n(s_n)$$

for every assignment $\{s_i\}_{i=1}^n$ to $\{\mathsf{X}_i\}_{i=1}^n$. Then the family $\{\mathsf{X}_i\}_{i=1}^n$ is mutually independent, and for each $i = 1, \ldots, n$, the distribution of X_i is f_i.

Proof. It suffices to show that the distribution of X_i is f_i for $i = 1, \ldots, n$. Using a calculation similar to that used in the proof of Theorem 3, one can eliminate one of the variables, say X_n, as follows:

$$\begin{aligned}
&\Pr[(\mathsf{X}_1 = s_1) \cap \cdots \cap (\mathsf{X}_{n-1} = s_{n-1})] \\
&\quad = \sum_{s_n} \Pr[(\mathsf{X}_1 = s_1) \cap \cdots \cap (\mathsf{X}_{n-1} = s_{n-1}) \cap (\mathsf{X}_n = s_n)] \\
&\quad = \sum_{s_n} f_1(s_1) \cdots f_{n-1}(s_{n-1}) f_n(s_n) \\
&\quad = f_1(s_1) \cdots f_{n-1}(s_{n-1}) \sum_{s_n} f_n(s_n) \\
&\quad = f_1(s_1) \cdots f_{n-1}(s_{n-1}).
\end{aligned}$$

We can repeat this calculation, eliminating $\mathsf{X}_{n-2}, \mathsf{X}_{n-3}$, etc., until we get $\Pr[\mathsf{X}_1 = s_1] = f_1(s_1)$. Similar calculations yield $\Pr[\mathsf{X}_i = s_i] = f_i(s_i)$ for $i = 1, \ldots, n$. □

Theorem 6 immediately implies the following:

Theorem 7. *Let $\{\mathsf{X}_i\}_{i=1}^n$ be a family of random variables, where each X_i takes values in a finite set S_i. Then the following are equivalent:*

(i) $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ *is uniformly distributed over* $S_1 \times \cdots \times S_n$*;*

(ii) $\{\mathsf{X}_i\}_{i=1}^n$ *is mutually independent and each* X_i *is uniformly distributed over* S_i*.*

In Theorem 7, the fact that (ii) implies (i) follows immediately from the definitions; it is in proving that (i) implies (ii) that we make use of Theorem 6.

Theorem 6 also immediately implies the following theorem, which can be used to synthesize independent random variables "out of thin air," by taking the product of appropriate probability distributions.

Theorem 8. *Suppose* $\Pr$ *is the product distribution* $\Pr_1 \cdots \Pr_n$*, where each* $\Pr_i$ *is a probability distribution on a sample space* Ω_i*, so that the sample space of* $\Pr$ *is* $\Omega = \Omega_1 \times \cdots \times \Omega_n$*. For each* $i = 1, \ldots, n$*, let* X_i *be the random variable that projects on the* i*th coordinate, so that* $\mathsf{X}_i(\omega_1, \ldots, \omega_n) = \omega_i$*. Then* $\{\mathsf{X}_i\}_{i=1}^n$ *is mutually independent, and for each* $i = 1, \ldots, n$*, the distribution of* X_i *is* $\Pr_i$*.*

Example 23. Theorem 8 immediately implies that in Example 17, the family of indicator variables $\{\mathsf{X}_i\}_{i=1}^n$ is mutually independent. □

The following theorem gives us yet another way to establish independence.

Theorem 9. *Suppose* $\{\mathsf{X}_i\}_{i=1}^n$ *is a mutually independent family of random variables. Further, suppose that for* $i = 1, \ldots, n$*, we have* $\mathsf{Y}_i = g_i(\mathsf{X}_i)$ *for some function* g_i*. Then* $\{\mathsf{Y}_i\}_{i=1}^n$ *is mutually independent.*

Proof. It suffices to prove the theorem for $n = 2$. The general case follows easily by induction, by considering the two random variables $(\mathsf{X}_1, \ldots, \mathsf{X}_{n-1})$ and X_n and applying Theorem 5.

For $i = 1, 2$, let t_i be any value in the image of Y_i, and let $S_i' := g_i^{-1}(\{t_i\})$. We have

$$\begin{aligned}
\Pr[(\mathsf{Y}_1 = t_1) \cap (\mathsf{Y}_2 = t_2)] &= \Pr\Big[\Big(\bigcup_{s_1 \in S_1'} (\mathsf{X}_1 = s_1)\Big) \cap \Big(\bigcup_{s_2 \in S_2'} (\mathsf{X}_2 = s_2)\Big)\Big] \\
&= \Pr\Big[\bigcup_{s_1 \in S_1'} \bigcup_{s_2 \in S_2'} \Big((\mathsf{X}_1 = s_1) \cap (\mathsf{X}_2 = s_2)\Big)\Big] \\
&= \sum_{s_1 \in S_1'} \sum_{s_2 \in S_2'} \Pr[(\mathsf{X}_1 = s_1) \cap (\mathsf{X}_2 = s_2)] \\
&= \sum_{s_1 \in S_1'} \sum_{s_2 \in S_2'} \Pr[\mathsf{X}_1 = s_1] \Pr[\mathsf{X}_2 = s_2] \\
&= \Big(\sum_{s_1 \in S_1'} \Pr[\mathsf{X}_1 = s_1]\Big)\Big(\sum_{s_2 \in S_2'} \Pr[\mathsf{X}_2 = s_2]\Big) \\
&= \Pr\Big[\bigcup_{s_1 \in S_1'} (\mathsf{X}_1 = s_1)\Big] \Pr\Big[\bigcup_{s_2 \in S_2'} (\mathsf{X}_2 = s_2)\Big] = \Pr[\mathsf{Y}_1 = t_1] \Pr[\mathsf{Y}_2 = t_2]. \quad \square
\end{aligned}$$

Example 24. As a special case of the previous theorem, suppose $\{\mathsf{X}_i\}_{i=1}^n$ is a mutually independent family of random variables, where each X_i takes values in some set S_i. Also suppose that $T_i \subseteq S_i$ for $i = 1, \ldots, n$, and we want to compute the probablity

$$\Pr[(\mathsf{X}_1 \in T_1) \cap \cdots \cap (\mathsf{X}_n \in T_n)].$$

For $i = 1, \ldots, n$, define the function $f_i : S_i \to \{0, 1\}$ to be 1 if its input lies in T_i and 0 otherwise. Applying the above theorem with these functions, we obtain

$$\begin{aligned}
\Pr[(\mathsf{X}_1 \in T_1) \cap \cdots \cap (\mathsf{X}_n \in T_n)] &= \Pr[(f_1(\mathsf{X}_1) = 1) \cap \cdots \cap (f_n(\mathsf{X}_n) = 1)] \\
&= \Pr[f_1(\mathsf{X}_1) = 1] \cdots \Pr[f_n(\mathsf{X}_n) = 1] \\
&= \Pr[\mathsf{X}_1 \in T_1] \cdots \Pr[\mathsf{X}_n \in T_n]. \quad \square
\end{aligned}$$

Example 25. We now develop an extremely useful technique for constructing k-wise independent families of random variables. Let p be a prime and let $Poly_k$ be the set of all polynomials in $\mathbb{Z}_p[\mathtt{X}]$ of degree less than k, that is, all polynomials of the form $g = \sum_{j=0}^{k-1} a_j \mathtt{X}^j \in \mathbb{Z}_p[\mathtt{X}]$. Now suppose we choose a polynomial at random from $Poly_k$. We can model this using a random variable G uniformly distributed over $Poly_k$. We can evaluate this random polynomial at any point $\gamma \in \mathbb{Z}_p$, which defines a new random variable $\mathsf{G}(\gamma)$. We claim that the family of random variables $\{\mathsf{G}(\gamma)\}_{\gamma \in \mathbb{Z}_p}$ is a k-wise independent family of random variables, with each $\mathsf{G}(\gamma)$ uniformly distributed over $\mathbb{Z}_p$.

Using Theorem 7, it will suffice to show that for any choice of distinct evaluation points $\gamma_1, \ldots, \gamma_k \in \mathbb{Z}_p$, the evaluation vector $\mathsf{Z} := (\mathsf{G}(\gamma_1), \ldots, \mathsf{G}(\gamma_k))$ is uniformly distributed over $\mathbb{Z}_p^k$. To prove this, we observe that Langrange's polynomial interpolation theorem says that the map

$$\begin{aligned} Eval: \quad & Poly_k \to \mathbb{Z}_p^k \\ & g \mapsto (g(\gamma_1), \ldots, g(\gamma_k)) \end{aligned}$$

is bijective. Since $\mathsf{Z} = Eval(\mathsf{G})$ and G is uniformly distributed over $Poly_k$, it follows that Z is also uniformly distributed over $\mathbb{Z}_p^k$.

Finally, we observe that $\{\mathsf{G}(\gamma)\}_{\gamma \in \mathbb{Z}_p}$ is not $(k+1)$-wise independent. This again follows from Lagrange interpolation: the values of a polynomial of degree less than k at k distinct evaluation points completely determine the coefficients of the polynomial, and hence its value at any other evaluation point. $\square$

Example 26. The previous example can be extended to illustrate the notion of **threshold secret sharing**. Suppose Alice has a secret that she wants to back up to some servers in "the cloud". The secret sharing scheme we present allows Alice to split her secret into some number, say n, of "shares". She stores each of these n shares on a different server. None of these shares by themselves reveal anything about her secret. In fact, we can arrange that for an arbitrary parameter $k < n$, no k of these shares taken together reveal anything about her secret, but any subset of $k+1$ shares can be used to reconstruct her secret. Such a scheme is called a $(k+1)$**-out-of-**n **secret sharing scheme**.

Here is how it works. As in the previous example, we work with a prime p, and assume that Alice's secret can be encoded as an element of $\mathbb{Z}_p$. We model Alice's secret as a random variable S with some arbitrary distribution over $\mathbb{Z}_p$. Then, as in the previous example, Alice chooses a random polynomial $\mathsf{G} \in \mathbb{Z}_p[\mathtt{X}]$ of degree less than k and sets $\mathsf{H} := \mathsf{G} + \mathsf{S}\mathtt{X}^k$. We insist that G and S are independent. Alice then evaluates H at n distinct evaluation points $\gamma_1, \ldots, \gamma_n \in \mathbb{Z}_p$, obtaining $\mathsf{S}_i := \mathsf{H}(\gamma_i) \in \mathbb{Z}_p$ for $i = 1, \ldots, n$. The values $\mathsf{S}_1, \ldots, \mathsf{S}_n$ are the shares of her secret.

First, observe that since H has degree at most k, from any $k+1$ shares, Alice can reconstruct H, and hence her secret.

Second, we claim that no subset of k (or fewer) shares reveals any useful information about her secret. To this end, let us fix k distinct evaluation points $\gamma_{i_1}, \ldots, \gamma_{i_k} \in \mathbb{Z}_p$. Our goal is to show that the family of random variables $\mathsf{S}, \mathsf{S}_{i_1}, \ldots, \mathsf{S}_{i_k}$ is mutually independent. Consider an arbitrary fixed secret $\sigma \in \mathbb{Z}_p$. By the previous example, we know that $(\mathsf{G}(\gamma_{i_1}), \ldots, \mathsf{G}(\gamma_{i_k}))$ is uniformly distributed over $\mathbb{Z}_p^k$. It follows that

$$(\mathsf{G}(\gamma_{i_1}) + \sigma\gamma_{i_1}^k, \ldots, \mathsf{G}(\gamma_{i_k}) + \sigma\gamma_{i_k}^k)$$

is also uniformly distributed over $\mathbb{Z}_p^k$, since that map that sends $(v_1, \ldots, v_k)$ to $(v_1 + \sigma\gamma_{i_1}^k, \ldots, v_1 + \sigma\gamma_{i_k}^k)$ is a bijection on $\mathbb{Z}_p^k$.

From the above, it follows that for any fixed $\sigma, \sigma_1, \ldots, \sigma_k \in \mathbb{Z}_p$, we have

$$\begin{aligned}
&\Pr[(\mathsf{S} = \sigma) \cap (\mathsf{S}_{i_1} = \sigma_1) \cap \cdots \cap (\mathsf{S}_{i_k} = \sigma_k)] \\
&\quad = \Pr[(\mathsf{S} = \sigma) \cap (\mathsf{G}(\gamma_{i_1}) + \sigma\gamma_{i_1}^k = \sigma_1) \cap \cdots \cap (\mathsf{G}(\gamma_{i_k}) + \sigma\gamma_{i_k}^k = \sigma_k)] \\
&\quad = \Pr[\mathsf{S} = \sigma] \Pr[(\mathsf{G}(\gamma_{i_1}) + \sigma\gamma_{i_1}^k = \sigma_1) \cap \cdots \cap (\mathsf{G}(\gamma_{i_k}) + \sigma\gamma_{i_k}^k = \sigma_k)] \\
&\qquad \text{(independence of } \mathsf{S} \text{ and } \mathsf{G}\text{, and Theorem 9)} \\
&\quad = \Pr[\mathsf{S} = \sigma](1/p^k).
\end{aligned}$$

Our claim then follows from Theorem 6.

Threshold secret sharing schemes have other, more serious applications. For example, *Time* magazine reported in the early 1990's that, at the time, Russia protected its nuclear launch codes using a 2-out-of-3 secret sharing scheme. □

4 Expectation and variance

Let Pr be a probability distribution on a sample space Ω. If X is a real-valued random variable, then its **expected value**, or **expectation**, is defined as

$$\mathrm{E}[\mathsf{X}] := \sum_{\omega \in \Omega} \mathsf{X}(\omega) \Pr(\omega). \tag{12}$$

If S is the image of X, and if for each $s \in S$ we group together the terms in (12) with $\mathsf{X}(\omega) = s$, then we see that

$$\mathrm{E}[\mathsf{X}] = \sum_{s \in S} s \Pr[\mathsf{X} = s]. \tag{13}$$

From (13), it is clear that $\mathrm{E}[\mathsf{X}]$ depends only on the distribution of X: if X' is another random variable with the same (or essentially the same) distribution as X, then $\mathrm{E}[\mathsf{X}] = \mathrm{E}[\mathsf{X}']$.

More generally, suppose X is an arbitrary random variable (not necessarily real valued) whose image is S, and f is a real-valued function on S. Then again, if for each $s \in S$ we group together the terms in (12) with $\mathsf{X}(\omega) = s$, we see that

$$\mathrm{E}[f(\mathsf{X})] = \sum_{s \in S} f(s) \Pr[\mathsf{X} = s]. \tag{14}$$

We make a few trivial observations about expectation, which the reader may easily verify. First, if X is equal to a constant c (i.e., $\mathsf{X}(\omega) = c$ for every $\omega \in \Omega$), then $\mathrm{E}[\mathsf{X}] = \mathrm{E}[c] = c$. Second, if X and Y are random variables such that $\mathsf{X} \geq \mathsf{Y}$ (i.e., $\mathsf{X}(\omega) \geq \mathsf{Y}(\omega)$ for every $\omega \in \Omega$), then $\mathrm{E}[\mathsf{X}] \geq \mathrm{E}[\mathsf{Y}]$. Similarly, if $\mathsf{X} > \mathsf{Y}$, then $\mathrm{E}[\mathsf{X}] > \mathrm{E}[\mathsf{Y}]$.

In calculating expectations, one rarely makes direct use of (12), (13), or (14), except in rather trivial situations. The next two theorems develop tools that are often quite effective in calculating expectations.

Theorem 10 (Linearity of expectation). *If X and Y are real-valued random variables, and a is a real number, then*

$$\mathrm{E}[\mathsf{X} + \mathsf{Y}] = \mathrm{E}[\mathsf{X}] + \mathrm{E}[\mathsf{Y}] \quad \textit{and} \quad \mathrm{E}[a\mathsf{X}] = a\,\mathrm{E}[\mathsf{X}].$$

Proof. It is easiest to prove this using the defining equation (12) for expectation. For $\omega \in \Omega$, the value of the random variable $\mathsf{X} + \mathsf{Y}$ at ω is by definition $\mathsf{X}(\omega) + \mathsf{Y}(\omega)$, and so we have

$$\begin{aligned}\mathrm{E}[\mathsf{X} + \mathsf{Y}] &= \sum_{\omega} (\mathsf{X}(\omega) + \mathsf{Y}(\omega)) \Pr(\omega) \\ &= \sum_{\omega} \mathsf{X}(\omega) \Pr(\omega) + \sum_{\omega} \mathsf{Y}(\omega) \Pr(\omega) \\ &= \mathrm{E}[\mathsf{X}] + \mathrm{E}[\mathsf{Y}].\end{aligned}$$

For the second part of the theorem, by a similar calculation, we have

$$\mathrm{E}[a\mathsf{X}] = \sum_{\omega} (a\mathsf{X}(\omega)) \Pr(\omega) = a \sum_{\omega} \mathsf{X}(\omega) \Pr(\omega) = a \,\mathrm{E}[\mathsf{X}]. \quad \square$$

More generally, the above theorem implies (using a simple induction argument) that if $\{\mathsf{X}_i\}_{i \in I}$ is a finite family of real-valued random variables, then we have

$$\mathrm{E}\Big[\sum_{i \in I} \mathsf{X}_i\Big] = \sum_{i \in I} \mathrm{E}[\mathsf{X}_i]. \tag{15}$$

So we see that expectation is linear; however, expectation is not in general multiplicative, except in the case of *independent* random variables:

Theorem 11. *If* X *and* Y *are independent, real-valued random variables, then* $\mathrm{E}[\mathsf{XY}] = \mathrm{E}[\mathsf{X}]\,\mathrm{E}[\mathsf{Y}]$.

Proof. It is easiest to prove this using (14), with the function $f(s,t) := st$ applied to the random variable (X, Y). We have

$$\begin{aligned}\mathrm{E}[\mathsf{XY}] &= \sum_{s,t} st \Pr[(\mathsf{X} = s) \cap (\mathsf{Y} = t)] \\ &= \sum_{s,t} st \Pr[\mathsf{X} = s] \Pr[\mathsf{Y} = t] \\ &= \Big(\sum_{s} s \Pr[\mathsf{X} = s]\Big)\Big(\sum_{t} t \Pr[\mathsf{Y} = t]\Big) \\ &= \mathrm{E}[\mathsf{X}]\,\mathrm{E}[\mathsf{Y}]. \quad \square\end{aligned}$$

More generally, the above theorem implies (using a simple induction argument) that if $\{\mathsf{X}_i\}_{i \in I}$ is a finite, mutually independent family of real-valued random variables, then

$$\mathrm{E}\Big[\prod_{i \in I} \mathsf{X}_i\Big] = \prod_{i \in I} \mathrm{E}[\mathsf{X}_i]. \tag{16}$$

The following simple facts are also sometimes quite useful in calculating expectations:

Theorem 12. *Let* X *be a* 0/1*-valued random variable. Then* $\mathrm{E}[\mathsf{X}] = \Pr[\mathsf{X} = 1]$.

Proof. $\mathrm{E}[\mathsf{X}] = 0 \cdot \Pr[\mathsf{X} = 0] + 1 \cdot \Pr[\mathsf{X} = 1] = \Pr[\mathsf{X} = 1]$. $\square$

Theorem 13 (Tail sum formula). *If* X *is a random variable that takes only non-negative integer values, then*

$$\mathrm{E}[\mathsf{X}] = \sum_{i \geq 1} \Pr[\mathsf{X} \geq i].$$

Note that since X has a finite image, the sum appearing above is finite.

Proof. Let $p_i := \Pr[\mathsf{X} = i]$ for $i = 1, 2, \ldots$, and consider the matrix

$$\begin{pmatrix} p_1 & & & \\ p_2 & p_2 & & \\ p_3 & p_3 & p_3 & \\ \vdots & \vdots & \vdots & \ldots \end{pmatrix}.$$

The ith row sums to $i \Pr[\mathsf{X} = i]$, and summing row by row, we see that the sum of all the numbers in the matrix is precisely $\mathrm{E}[\mathsf{X}]$. However, notice that the ith column sums to $\Pr[\mathsf{X} \geq i]$. Therefore, summing column by column, we see that the sum of all the numbers in the matrix is also equal to $\sum_{i \geq 1} \Pr[\mathsf{X} \geq i]$. $\square$

Example 27. Let X be uniformly distributed over $\{1, \ldots, m\}$. Let us compute $\mathrm{E}[\mathsf{X}]$. We have

$$\mathrm{E}[\mathsf{X}] = \sum_{s=1}^{m} s \cdot \frac{1}{m} = \frac{m(m+1)}{2} \cdot \frac{1}{m} = \frac{m+1}{2}. \quad \square$$

Example 28. As a special case of the previous example, the expected value of a roll of a die is $(6+1)/2 = 3.5$. $\square$

Example 29. Let X be a random variable with a binomial distribution, as in Example 17, that counts the number of successes among n Bernoulli trials, each of which succeeds with probability p. Let us compute $\mathrm{E}[\mathsf{X}]$. We can write X as the sum of indicator variables, $\mathsf{X} = \sum_{i=1}^{n} \mathsf{X}_i$, where X_i is the indicator variable for the event that the ith trial succeeds; each X_i takes the value 1 with probability p and 0 with probability $q := 1 - p$. By Theorem 12, we have $\mathrm{E}[\mathsf{X}_i] = p$, for $i = 1, \ldots, n$. By linearity of expectation, we have

$$\mathrm{E}[\mathsf{X}] = \sum_{i=1}^{n} \mathrm{E}[\mathsf{X}_i] = np. \quad \square$$

Example 30. Let $\mathsf{X} := \sum_{i=1}^{n} \mathsf{X}_i$, where the family random variables $\mathsf{X}_1, \ldots, \mathsf{X}_n$ is mutually independent, and each X_i is uniformly distributed over $\{\pm 1\}$. Let us compute $\mathrm{E}[\mathsf{X}]$ and $\mathrm{E}[\mathsf{X}^2]$. Observe that for each $i = 1, \ldots, n$, we have $\mathrm{E}[\mathsf{X}_i] = 1 \cdot (1/2) + (-1) \cdot (1/2) = 0$. Therefore, by linearity of expectation, we have

$$\mathrm{E}[\mathsf{X}] = \sum_{i=1}^{n} \mathrm{E}[\mathsf{X}_i] = 0.$$

To compute $\mathrm{E}[\mathsf{X}^2]$, we first observe that we can write

$$\begin{aligned} \mathsf{X}^2 &= \Big(\sum_{i=1}^{n} \mathsf{X}_i\Big)\Big(\sum_{j=1}^{n} \mathsf{X}_j\Big) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathsf{X}_i \mathsf{X}_j \\ &= \sum_{i=1}^{n} \mathsf{X}_i^2 + \sum_{i \neq j} \mathsf{X}_i \mathsf{X}_j, \end{aligned}$$

where the sum $\sum_{i \neq j} \mathsf{X}_i \mathsf{X}_j$ is over all $n(n-1)$ pairs of indices (i, j) with $i \neq j$.

By linearity of expectation, we have

$$\mathrm{E}[\mathsf{X}^2] = \sum_{i=1}^{n} \mathrm{E}[\mathsf{X}_i^2] + \sum_{i \neq j} \mathrm{E}[\mathsf{X}_i \mathsf{X}_j]. \tag{17}$$

Note that each term in the first sum is 1, since $\mathrm{E}[\mathsf{X}_i^2] = (1)^2 \cdot (1/2) + (-1)^2 \cdot (1/2) = 1$. As there are n such terms, the first sum in (17) is n. For the terms in the second sum in (17), notice that since X_i and X_j are independent, Theorem 11 says that $\mathrm{E}[\mathsf{X}_i \mathsf{X}_j] = \mathrm{E}[\mathsf{X}_i]\,\mathrm{E}[\mathsf{X}_j] = 0$. Thus, the second sum in (17) is zero. Therefore, $\mathrm{E}[\mathsf{X}^2] = n$. □

Example 31. Suppose we roll four dice. For $i = 1, \ldots, 4$, let X_i be the value of the ith die. So $\mathsf{X}_1, \ldots, \mathsf{X}_4$ is a mutually independent family of random variables, where each X_i is uniformly distributed over $\{1, \ldots, 6\}$. Let M be the minimum number showing on any of the dice; that is, $\mathsf{M} := \min(\mathsf{X}_1, \ldots, \mathsf{X}_4)$. Let us compute $\mathrm{E}[\mathsf{M}]$ using the tail sum formula (Theorem 13), which says that

$$\mathrm{E}[\mathsf{M}] = \sum_{j=1}^{6} \Pr[\mathsf{M} \geq j].$$

Now, $\mathsf{M} \geq j$ occurs if and only if $\mathsf{X}_i \geq j$ for all $i = 1, \ldots, 4$. For each $i = 1, \ldots, 4$, we have $\Pr[\mathsf{X}_i \geq j] = (7 - j)/6$. By independence (and Theorem 9), we have

$$\Pr[\mathsf{M} \geq j] = \Pr[(\mathsf{X}_1 \geq j) \cap \cdots \cap (\mathsf{X}_4 \geq j)] = \Pr[\mathsf{X}_1 \geq j] \cdots \Pr[\mathsf{X}_4 \geq j] = \Big(\frac{7 - j}{6}\Big)^4.$$

So we have

$$\mathrm{E}[\mathsf{M}] = \sum_{j=1}^{6} \Pr[\mathsf{M} \geq j] = \sum_{j=1}^{6} \Big(\frac{7 - j}{6}\Big)^4 = \frac{6^4 + 5^4 + 4^4 + 3^4 + 2^4 + 1^4}{6^4} \approx 1.75.$$

Compare this to the expected value of a single roll of a die, which by Example 28 is 3.5. □

Example 32. Continuing with the previous example, let us compute the expected value of the sum S of the three largest dice. We can express S as

$$\mathsf{S} = \mathsf{X}_1 + \cdots + \mathsf{X}_4 - \mathsf{M},$$

and by linearity of expectation, we calculate

$$\mathrm{E}[\mathsf{S}] = 4 \cdot 3.5 - \mathrm{E}[\mathsf{M}] = 14 - \mathrm{E}[\mathsf{M}] \approx 12.25. \quad \square$$

Let $\mathcal{B}$ be an event with $\Pr[\mathcal{B}] \neq 0$, and let X be a real-valued random variable. We define the **conditional expectation of X given $\mathcal{B}$**, denoted $\mathrm{E}[\mathsf{X} \mid \mathcal{B}]$, to be the expected value of the X *relative to the conditional distribution* $\Pr(\cdot \mid \mathcal{B})$, so that

$$\mathrm{E}[\mathsf{X} \mid \mathcal{B}] = \sum_{\omega \in \Omega} \mathsf{X}(\omega) \Pr(\omega \mid \mathcal{B}) = \Pr[\mathcal{B}]^{-1} \sum_{\omega \in \mathcal{B}} \mathsf{X}(\omega) \Pr(\omega).$$

Analogous to (13), if S is the image of X, we have

$$\mathrm{E}[\mathsf{X} \mid \mathcal{B}] = \sum_{s \in S} s \Pr[\mathsf{X} = s \mid \mathcal{B}]. \tag{18}$$

Furthermore, suppose I is a finite index set, and $\{\mathcal{B}_i\}_{i\in I}$ is a finite, pairwise disjoint family of events whose union is Ω. If for each $i \in I$ we group together the terms in (12) with $\omega \in \mathcal{B}_i$, we obtain the **law of total expectation**:

$$\mathrm{E}[\mathsf{X}] = \sum_{i\in I} \mathrm{E}[\mathsf{X} \mid \mathcal{B}_i] \Pr[\mathcal{B}_i], \tag{19}$$

with the understanding that if any of the $\mathcal{B}_i$'s occur with zero probability, the corresponding terms in (19) are excluded.

Example 33. Let X denote the value of a roll of a die. Let $\mathcal{A}$ be the event that X is even. Then the conditional distribution of X given $\mathcal{A}$ is essentially the uniform distribution on $\{2, 4, 6\}$, and hence

$$\mathrm{E}[\mathsf{X} \mid \mathcal{A}] = \frac{2+4+6}{3} = 4.$$

Similarly, the conditional distribution of X given $\overline{\mathcal{A}}$ is essentially the uniform distribution on $\{1, 3, 5\}$, and so

$$\mathrm{E}[\mathsf{X} \mid \overline{\mathcal{A}}] = \frac{1+3+5}{3} = 3.$$

Using the law of total expectation, we can compute the expected value of X as follows:

$$\mathrm{E}[\mathsf{X}] = \mathrm{E}[\mathsf{X} \mid \mathcal{A}] \Pr[\mathcal{A}] + \mathrm{E}[\mathsf{X} \mid \overline{\mathcal{A}}] \Pr[\overline{\mathcal{A}}] = 4 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = \frac{7}{2},$$

which agrees with the calculation in Example 28. □

Example 34. Suppose we toss a coin, and set $\mathsf{B} := 0$ if it comes up heads and set $\mathsf{B} := 1$ if it comes up tails. We also roll two dice. Let X and Y be their values. So we are assuming that B is uniformly distributed over $\{0, 1\}$, X and Y are uniformly distributed over $\{1, \ldots, 6\}$, and that B, X, and Y are mutually independent.

If $\mathsf{B} = 0$, set $\mathsf{Z} = \mathsf{X} + \mathsf{Y}$; if $\mathsf{B} = 1$, set $\mathsf{Z} = \mathsf{XY}$. We want to compute $\mathrm{E}[\mathsf{Z}]$. So we can use the law of total expectation to treat the two cases $\mathsf{B} = 0$ and $\mathsf{B} = 1$ separately. We have

$$\begin{aligned}
\mathrm{E}[\mathsf{Z}] &= \mathrm{E}[\mathsf{Z} \mid \mathsf{B} = 0] \Pr[\mathsf{B} = 0] + \mathrm{E}[\mathsf{Z} \mid \mathsf{B} = 1] \Pr[\mathsf{B} = 1] \\
&= \mathrm{E}[\mathsf{X} + \mathsf{Y} \mid \mathsf{B} = 0] \cdot \tfrac{1}{2} + \mathrm{E}[\mathsf{XY} \mid \mathsf{B} = 1] \cdot \tfrac{1}{2} \\
&= \mathrm{E}[\mathsf{X} + \mathsf{Y}] \cdot \tfrac{1}{2} + \mathrm{E}[\mathsf{XY}] \cdot \tfrac{1}{2} \qquad \text{(by mutual independence of } \mathsf{B}, \mathsf{X}, \text{ and } \mathsf{Y}) \\
&= (\mathrm{E}[\mathsf{X}] + \mathrm{E}[\mathsf{Y}]) \cdot \tfrac{1}{2} + \mathrm{E}[\mathsf{X}]\,\mathrm{E}[\mathsf{Y}] \cdot \tfrac{1}{2} \quad \text{(by independence of } \mathsf{X} \text{ and } \mathsf{Y}\text{, and Theorems 10 and 11)} \\
&= 2 \cdot \tfrac{7}{2} \cdot \tfrac{1}{2} + \left(\tfrac{7}{2}\right)^2 \cdot \tfrac{1}{2} \ .
\end{aligned}$$

In the above calculation, we exploited the fact that the distribution of $\mathsf{X} + \mathsf{Y}$ given $\mathsf{B} = 0$ is the same as the distribution of $\mathsf{X} + \mathsf{Y}$, which follows from the independence of $(\mathsf{X} + \mathsf{Y})$ and B, and implies that $\mathrm{E}[\mathsf{X} + \mathsf{Y} \mid \mathsf{B} = 0] = \mathrm{E}[\mathsf{X} + \mathsf{Y}]$. Similarly, $\mathrm{E}[\mathsf{XY} \mid \mathsf{B} = 1] = \mathrm{E}[\mathsf{XY}]$. □

Let X be a real-valued random variable with $\mu := \mathrm{E}[\mathsf{X}]$. The **variance** of X is $\mathrm{Var}[\mathsf{X}] := \mathrm{E}[(\mathsf{X} - \mu)^2]$. The variance provides a measure of the spread or dispersion of the distribution of X around its expected value. Note that since $(\mathsf{X} - \mu)^2$ takes only non-negative values, variance is always non-negative.

Theorem 14. *Let X be a real-valued random variable, with $\mu := \mathrm{E}[\mathsf{X}]$, and let a and b be real numbers. Then we have*

(i) $\mathrm{Var}[\mathsf{X}] = \mathrm{E}[\mathsf{X}^2] - \mu^2$,

(ii) $\mathrm{Var}[a\mathsf{X}] = a^2 \mathrm{Var}[\mathsf{X}]$, *and*

(iii) $\mathrm{Var}[\mathsf{X} + b] = \mathrm{Var}[\mathsf{X}]$.

Proof. For part (i), observe that

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

where in the third equality, we used the fact that expectation is linear, and in the fourth equality, we used the fact that $\mathrm{E}[c] = c$ for constant c (in this case, $c = \mu^2$).

For part (ii), observe that

$$\begin{aligned}\mathrm{Var}[a\mathsf{X}] &= \mathrm{E}[a^2\mathsf{X}^2] - \mathrm{E}[a\mathsf{X}]^2 = a^2\,\mathrm{E}[\mathsf{X}^2] - (a\mu)^2 \\ &= a^2(\mathrm{E}[\mathsf{X}^2] - \mu^2) = a^2\,\mathrm{Var}[\mathsf{X}],\end{aligned}$$

where we used part (i) in the first and fourth equality, and the linearity of expectation in the second.

Part (iii) follows by a similar calculation:

$$\begin{aligned}\mathrm{Var}[\mathsf{X} + b] &= \mathrm{E}[(\mathsf{X} + b)^2] - (\mu + b)^2 \\ &= (\mathrm{E}[\mathsf{X}^2] + 2b\mu + b^2) - (\mu^2 + 2b\mu + b^2) \\ &= \mathrm{E}[\mathsf{X}^2] - \mu^2 = \mathrm{Var}[\mathsf{X}]. \quad \square\end{aligned}$$

The following is an immediate consequence of part (i) of Theorem 14, and the fact that variance is always non-negative:

Theorem 15. *If* X *is a real-valued random variable, then* $\mathrm{E}[\mathsf{X}^2] \geq \mathrm{E}[\mathsf{X}]^2$.

Theorem 15 is a special case of **Jensen's inequality**, which says that if f is convex on an interval (see §A6), and X is a random variable taking values in that interval, then $\mathrm{E}[f(\mathsf{X})] \geq f(\mathrm{E}[\mathsf{X}])$. It is not hard to prove Jensen's inequality by induction on the size of the image of X — we leave that as an exercise to the reader.

Unlike expectation, the variance of a sum of random variables is not equal to the sum of the variances, unless the variables are *pairwise independent*:

Theorem 16. *If* $\{\mathsf{X}_i\}_{i \in I}$ *is a finite, pairwise independent family of real-valued random variables, then*

$$\mathrm{Var}\Big[\sum_{i \in I} \mathsf{X}_i\Big] = \sum_{i \in I} \mathrm{Var}[\mathsf{X}_i].$$

Proof. We have

$$\begin{aligned}\mathrm{Var}\Big[\sum_{i\in I}\mathsf{X}_i\Big] &= \mathrm{E}\Big[\Big(\sum_{i\in I}\mathsf{X}_i\Big)^2\Big] - \Big(\mathrm{E}\Big[\sum_{i\in I}\mathsf{X}_i\Big]\Big)^2 \\ &= \sum_{i\in I}\mathrm{E}[\mathsf{X}_i^2] + \sum_{\substack{i,j\in I\\ i\neq j}}\big(\mathrm{E}[\mathsf{X}_i\mathsf{X}_j] - \mathrm{E}[\mathsf{X}_i]\,\mathrm{E}[\mathsf{X}_j]\big) - \sum_{i\in I}\mathrm{E}[\mathsf{X}_i]^2 \\ &\quad \text{(by linearity of expectation and rearranging terms)} \\ &= \sum_{i\in I}\mathrm{E}[\mathsf{X}_i^2] - \sum_{i\in I}\mathrm{E}[\mathsf{X}_i]^2 \\ &\quad \text{(by pairwise independence and Theorem 11)} \\ &= \sum_{i\in I}\mathrm{Var}[\mathsf{X}_i]. \ \square\end{aligned}$$

Corresponding to Theorem 12, we have:

Theorem 17. *Let* X *be a* 0/1*-valued random variable, with* $p := \Pr[\mathsf{X} = 1]$ *and* $q := \Pr[\mathsf{X} = 0] = 1 - p$*. Then* $\mathrm{Var}[\mathsf{X}] = pq$*.*

Proof. We have $\mathrm{E}[\mathsf{X}] = p$ and $\mathrm{E}[\mathsf{X}^2] = \Pr[\mathsf{X}^2 = 1] = \Pr[\mathsf{X} = 1] = p$. Therefore,

$$\mathrm{Var}[\mathsf{X}] = \mathrm{E}[\mathsf{X}^2] - \mathrm{E}[\mathsf{X}]^2 = p - p^2 = p(1-p) = pq. \ \square$$

Example 35. Let X be uniformly distributed over $\{1,\ldots,m\}$. Let us compute $\mathrm{Var}[\mathsf{X}]$. As we calculated in Example 27, we have

$$\mathrm{E}[\mathsf{X}] = \frac{m+1}{2}.$$

We also have

$$\mathrm{E}[\mathsf{X}^2] = \sum_{s=1}^{m} s^2 \cdot \frac{1}{m} = \frac{m(m+1)(2m+1)}{6}\cdot\frac{1}{m} = \frac{(m+1)(2m+1)}{6}.$$

Therefore,

$$\mathrm{Var}[\mathsf{X}] = \mathrm{E}[\mathsf{X}^2] - \mathrm{E}[\mathsf{X}]^2 = \frac{m^2-1}{12}. \ \square$$

Example 36. Let X be a random variable with a binomial distribution, as in Example 17, that counts the number of successes among n Bernoulli trials, each of which succeeds with probability p. Let us compute and $\mathrm{Var}[\mathsf{X}]$. As in Example 29, we can write X as the sum of indicator variables, $\mathsf{X} = \sum_{i=1}^{n}\mathsf{X}_i$, where X_i is the indicator variable for the event that the ith trial succeeds; each X_i takes the value 1 with probability p and 0 with probability $q := 1 - p$. By Theorem 17, we have and $\mathrm{Var}[\mathsf{X}_i] = pq$ for $i = 1,\ldots,n$. The family of random variables $\{\mathsf{X}_i\}_{i=1}^{n}$ is mutually independent (see Example 23), and hence pairwise independent. By Theorem 16, we therefore have

$$\mathrm{Var}[\mathsf{X}] = \sum_{i=1}^{n}\mathrm{Var}[\mathsf{X}_i] = npq. \ \square$$

Example 37. Consider again Example 30, where $\mathsf{X} := \sum_{i=1}^{n}\mathsf{X}_i$, and the family random variables $\mathsf{X}_1,\ldots,\mathsf{X}_n$ is mutually independent, with each X_i is uniformly distributed over $\{\pm 1\}$. We can think of X as representing a "random walk": a drunken sailor takes n steps, where each step is either one

step to the right or one step to the left. We may ask: what is the expected distance of the sailor from his starting point after taking a random walk of n steps? So we want to estimate $\mathrm{E}[\mathsf{Y}]$, where $\mathsf{Y} := |\mathsf{X}|$. Theorem 15 says that $\mathrm{E}[\mathsf{Y}]^2 \leq \mathrm{E}[\mathsf{Y}^2] = \mathrm{E}[\mathsf{X}^2]$. We calculated $\mathrm{E}[\mathsf{X}^2] = n$ in Example 30. Therefore, $\mathrm{E}[\mathsf{Y}] \leq \sqrt{n}$. This gives us an upper bound of $\sqrt{n}$ on the expected distance from the starting point. It is also possible to compute a lower bound of $c\sqrt{n}$ for constant c, but we will not do this here. □

5 Some useful bounds

In this section, we present several theorems that can be used to bound the probability that a random variable deviates from its expected value by some specified amount.

Theorem 18 (Markov's inequality). *Let X be a random variable that takes only non-negative real values. Then for every $\alpha > 0$, we have*

$$\Pr[\mathsf{X} \geq \alpha] \leq \mathrm{E}[\mathsf{X}]/\alpha.$$

Proof. We have

$$\mathrm{E}[\mathsf{X}] = \sum_s s \Pr[\mathsf{X} = s] \;=\; \sum_{s<\alpha} s \Pr[\mathsf{X} = s] + \sum_{s \geq \alpha} s \Pr[\mathsf{X} = s],$$

where the summations are over elements s in the image of X. Since X takes only non-negative values, all of the terms are non-negative. Therefore,

$$\mathrm{E}[\mathsf{X}] \geq \sum_{s \geq \alpha} s \Pr[\mathsf{X} = s] \;\geq\; \sum_{s \geq \alpha} \alpha \Pr[\mathsf{X} = s] \;=\; \alpha \Pr[\mathsf{X} \geq \alpha]. \;\square$$

Markov's inequality may be the only game in town when nothing more about the distribution of X is known besides its expected value. However, if the variance of X is also known, then one can get a better bound.

Theorem 19 (Chebyshev's inequality). *Let X be a real-valued random variable, with $\mu := \mathrm{E}[\mathsf{X}]$ and $\nu := \mathrm{Var}[\mathsf{X}]$. Then for every $\alpha > 0$, we have*

$$\Pr[|\mathsf{X} - \mu| \geq \alpha] \leq \nu/\alpha^2.$$

Proof. Let $\mathsf{Y} := (\mathsf{X} - \mu)^2$. Then Y is always non-negative, and $\mathrm{E}[\mathsf{Y}] = \nu$. Applying Markov's inequality to Y, we have

$$\Pr[|\mathsf{X} - \mu| \geq \alpha] = \Pr[\mathsf{Y} \geq \alpha^2] \leq \nu/\alpha^2. \;\square$$

An important special case of Chebyshev's inequality is the following. Suppose that $\{\mathsf{X}_i\}_{i \in I}$ is a finite, non-empty, pairwise independent family of real-valued random variables, each with the same distribution. Let μ be the common value of $\mathrm{E}[\mathsf{X}_i]$, ν be the common value of $\mathrm{Var}[\mathsf{X}_i]$, and $n := |I|$. Set

$$\overline{\mathsf{X}} := \frac{1}{n} \sum_{i \in I} \mathsf{X}_i.$$

The variable $\overline{\mathsf{X}}$ is called the **sample mean** of $\{\mathsf{X}_i\}_{i \in I}$. By the linearity of expectation, we have $\mathrm{E}[\overline{\mathsf{X}}] = \mu$, and since $\{\mathsf{X}_i\}_{i \in I}$ is pairwise independent, it follows from Theorem 16 (along with part (ii) of Theorem 14) that $\mathrm{Var}[\overline{\mathsf{X}}] = \nu/n$. Applying Chebyshev's inequality, for every $\epsilon > 0$, we have

$$\Pr[|\overline{\mathsf{X}} - \mu| \geq \epsilon] \leq \frac{\nu}{n\epsilon^2}. \tag{20}$$

The inequality (20) says that for all $\epsilon > 0$, and for all $\delta > 0$, there exists n_0 (depending on ϵ and δ, as well as the variance ν) such that $n \geq n_0$ implies

$$\Pr[|\overline{\mathsf{X}} - \mu| \geq \epsilon] \leq \delta. \tag{21}$$

In words:

> *As n gets large, the sample mean closely approximates the expected value μ with high probability.*

This fact, known as the **law of large numbers**, justifies the usual intuitive interpretation given to expectation.

We now examine an even more specialized case of the above situation, where each X_i is takes values in some finite interval of length Δ.

We state the following theorem without proof.

Theorem 20 (Hoeffding inequality). *Let $\{\mathsf{X}_i\}_{i \in I}$ be a finite, non-empty, and mutually independent family of random variables, such that each X_i has expected value μ and takes values in an interval of length Δ. Also, let $n := |I|$ and $\overline{\mathsf{X}}$ be the sample mean of $\{\mathsf{X}_i\}_{i \in I}$. Then for every $\epsilon > 0$, we have:*

$$\Pr[|\overline{X} - \mu| \geq \epsilon] \leq 2e^{-2n\epsilon^2/\Delta^2}$$

Note that this theorem does not require that the X_i's have the same distribution. It generally gives very good bounds, typically much better than can be obtained from (20).

Example 38. Suppose we toss a fair coin 1000 times. Let X be the number of heads. We have $\mathrm{E}[\mathsf{X}] = 500$. Hoeffing's inequality, with $\mu = 1/2$, $\Delta = 1$, $n = 1000$, and $\epsilon = 1/10$ says:

$$\Pr[|\mathsf{X} - 500| \geq 100] = \Pr[|\mathsf{X}/1000 - 1/2| \geq 1/10] \leq 2e^{-2000(1/10)^2} = 2e^{-20} \approx 10^{-8.4}.$$

Note that this is a much better bound than what we get from (20). Indeed, by Theorem 17, the variance of each coin toss is $1/4$, and plugging that into (20), we get

$$\Pr[|\mathsf{X} - 500| \geq 100] = \Pr[|\mathsf{X}/1000 - 1/2| \geq 1/10] \leq \frac{1/4}{1000 \cdot (1/10)^2} = 1/40. \quad \square$$

6 Discrete probability distributions

In addition to working with probability distributions over finite sample spaces, one can also work with distributions over infinite sample spaces. If the sample space is *countable*, that is, either finite or *countably infinite* (see §A3), then the distribution is called a **discrete probability distribution**. We shall not consider any other types of probability distributions in this text. The theory developed in §§1–5 extends fairly easily to the countably infinite setting, and in this section, we discuss how this is done.

6.1 Basic definitions

To say that the sample space Ω is countably infinite simply means that there is a bijection f from the set of positive integers onto Ω; thus, we can enumerate the elements of Ω as $\omega_1, \omega_2, \omega_3, \ldots$, where $\omega_i := f(i)$.

As in the finite case, a **probability distribution on** Ω is a function $\Pr : \Omega \to [0, 1]$, where all the probabilities sum to 1, which means that the infinite series $\sum_{i=1}^{\infty} \Pr(\omega_i)$ converges to one.

Luckily, the convergence properties of an infinite series whose terms are all non-negative is invariant under a reordering of terms (see §A4), so it does not matter how we enumerate the elements of Ω.

Example 39. Suppose we toss a fair coin repeatedly until it comes up *heads*, and let k be the total number of tosses. We can model this experiment as a discrete probability distribution Pr, where the sample space consists of the set of all positive integers: for each positive integer k, $\Pr(k) := 2^{-k}$. We can check that indeed $\sum_{k=1}^{\infty} 2^{-k} = 1$, as required.

One may be tempted to model this experiment by setting up a probability distribution on the sample space of all infinite sequences of coin tosses; however, this sample space is not countably infinite, and so we cannot construct a discrete probability distribution on this space. While it is possible to extend the notion of a probability distribution to such spaces, this would take us too far afield. □

Example 40. More generally, suppose we repeatedly execute a Bernoulli trial until it succeeds, where each execution succeeds with probability $p > 0$ independently of the previous trials, and let k be the total number of trials executed. Then we associate the probability $\Pr(k) := q^{k-1}p$ with each positive integer k, where $q := 1 - p$, since we have $k - 1$ failures before the one success. One can easily check that these probabilities sum to 1. Such a distribution is called a **geometric distribution**. □

Example 41. The series $\sum_{k=1}^{\infty} 1/k^3$ converges to some positive number c. Therefore, we can define a probability distribution on the set of positive integers, where we associate with each $k \geq 1$ the probability $1/ck^3$. □

As in the finite case, an event is an arbitrary subset $\mathcal{A}$ of Ω. The probability $\Pr[\mathcal{A}]$ of $\mathcal{A}$ is defined as the sum of the probabilities associated with the elements of $\mathcal{A}$. This sum is treated as an infinite series when $\mathcal{A}$ is infinite. This series is guaranteed to converge, and its value does not depend on the particular enumeration of the elements of $\mathcal{A}$.

Example 42. Consider the geometric distribution discussed in Example 40, where p is the success probability of each Bernoulli trial, and $q := 1 - p$. For a given integer $i \geq 1$, consider the event $\mathcal{A}$ that the number of trials executed is at least i. Formally, $\mathcal{A}$ is the set of all integers greater than or equal to i. Intuitively, $\Pr[\mathcal{A}]$ should be q^{i-1}, since we perform at least i trials if and only if the first $i - 1$ trials fail. Just to be sure, we can compute

$$\Pr[\mathcal{A}] = \sum_{k \geq i} \Pr(k) = \sum_{k \geq i} q^{k-1}p = q^{i-1}p \sum_{k \geq 0} q^k = q^{i-1}p \cdot \frac{1}{1-q} = q^{i-1}. \quad \square$$

It is an easy matter to check that all the statements and theorems in §1 carry over *verbatim* to the case of countably infinite sample spaces. Moreover, Boole's inequality (6) and equality (7) are also valid for countably infinite families of events:

Theorem 21. *Suppose $\mathcal{A} := \bigcup_{i=1}^{\infty} \mathcal{A}_i$, where $\{\mathcal{A}_i\}_{i=1}^{\infty}$ is an infinite sequence of events. Then*

(i) $\Pr[\mathcal{A}] \leq \sum_{i=1}^{\infty} \Pr[\mathcal{A}_i]$, *and*

(ii) $\Pr[\mathcal{A}] = \sum_{i=1}^{\infty} \Pr[\mathcal{A}_i]$ *if* $\{\mathcal{A}_i\}_{i=1}^{\infty}$ *is pairwise disjoint.*

Proof. For $\omega \in \Omega$ and $\mathcal{B} \subseteq \Omega$, define $\delta_\omega[\mathcal{B}] := 1$ if $\omega \in \mathcal{B}$, and $\delta_\omega[\mathcal{B}] := 0$ if $\omega \notin \mathcal{B}$. First, suppose

that $\{\mathcal{A}_i\}_{i=1}^{\infty}$ is pairwise disjoint. Evidently, $\delta_\omega[\mathcal{A}] = \sum_{i=1}^{\infty} \delta_\omega[\mathcal{A}_i]$ for each $\omega \in \Omega$, and so

$$\begin{aligned}\Pr[\mathcal{A}] &= \sum_{\omega\in\Omega} \Pr(\omega)\delta_\omega[\mathcal{A}] = \sum_{\omega\in\Omega} \Pr(\omega) \sum_{i=1}^{\infty} \delta_\omega[\mathcal{A}_i] \\ &= \sum_{i=1}^{\infty} \sum_{\omega\in\Omega} \Pr(\omega)\delta_\omega[\mathcal{A}_i] = \sum_{i=1}^{\infty} \Pr[\mathcal{A}_i],\end{aligned}$$

where we use the fact that we may reverse the order of summation in an infinite double summation of non-negative terms (see §A5). That proves (ii), and (i) follows from (ii), applied to the sequence $\{\mathcal{A}'_i\}_{i=1}^{\infty}$, where $\mathcal{A}'_i := \mathcal{A}_i \setminus \bigcup_{j=1}^{i-1} \mathcal{A}_i$, as $\Pr[\mathcal{A}] = \sum_{i=1}^{\infty} \Pr[\mathcal{A}'_i] \leq \sum_{i=1}^{\infty} \Pr[\mathcal{A}_i]$. □

6.2 Conditional probability and independence

All of the definitions and results in §2 carry over *verbatim* to the countably infinite case. The law of total probability (equations (8) and (9)), as well as Bayes' theorem (10), extend to families of events $\{\mathcal{B}_i\}_{i\in I}$ indexed by any countably infinite set I.

6.3 Random variables

All of the definitions and results in §3 carry over *verbatim* to the countably infinite case. Note that the image of a random variable may be either finite or countably infinite. The definitions of independent families of random variables (k-wise and mutually) extend *verbatim* to infinite families.

6.4 Expectation and variance

We define the expected value of a real-valued random variable X exactly as in (12); that is, $\mathrm{E}[\mathsf{X}] := \sum_\omega \mathsf{X}(\omega)\Pr(\omega)$, but where this sum is now an infinite series. If this series converges absolutely (see §A4), then we say that X has **finite expectation**, or that $\mathrm{E}[\mathsf{X}]$ is **finite**. In this case, the series defining $\mathrm{E}[\mathsf{X}]$ converges to the same finite limit, regardless of the ordering of the terms.

If $\mathrm{E}[\mathsf{X}]$ is not finite, then under the right conditions, $\mathrm{E}[\mathsf{X}]$ may still exist, although its value will be $\pm\infty$. Consider first the case where X takes only non-negative values. In this case, if $\mathrm{E}[\mathsf{X}]$ is not finite, then we naturally define $\mathrm{E}[\mathsf{X}] := \infty$, as the series defining $\mathrm{E}[\mathsf{X}]$ diverges to ∞, regardless of the ordering of the terms. In the general case, we may define random variables X^+ and X^-, where

$$\mathsf{X}^+(\omega) := \max\{0, \mathsf{X}(\omega)\} \quad \text{and} \quad \mathsf{X}^-(\omega) := \max\{0, -\mathsf{X}(\omega)\},$$

so that $\mathsf{X} = \mathsf{X}^+ - \mathsf{X}^-$, and both X^+ and X^- take only non-negative values. Clearly, X has finite expectation if and only if both X^+ and X^- have finite expectation. Now suppose that $\mathrm{E}[\mathsf{X}]$ is not finite, so that one of $\mathrm{E}[\mathsf{X}^+]$ or $\mathrm{E}[\mathsf{X}^-]$ is infinite. If $\mathrm{E}[\mathsf{X}^+] = \mathrm{E}[\mathsf{X}^-] = \infty$, then we say that $\mathrm{E}[\mathsf{X}]$ **does not exist**; otherwise, we define $\mathrm{E}[\mathsf{X}] := \mathrm{E}[\mathsf{X}^+] - \mathrm{E}[\mathsf{X}^-]$, which is $\pm\infty$; in this case, the series defining $\mathrm{E}[\mathsf{X}]$ diverges to $\pm\infty$, regardless of the ordering of the terms.

Example 43. Let X be a random variable whose distribution is as in Example 41. Since the series $\sum_{k=1}^{\infty} 1/k^2$ converges and the series $\sum_{k=1}^{\infty} 1/k$ diverges, the expectation $\mathrm{E}[\mathsf{X}]$ is finite, while $\mathrm{E}[\mathsf{X}^2] = \infty$. One may also verify that the random variable $(-1)^{\mathsf{X}}\mathsf{X}^2$ has no expectation. □

All of the results in §4 carry over essentially unchanged, although one must pay some attention to "convergence issues."

If $\mathrm{E}[\mathsf{X}]$ exists, then we can regroup the terms in the series $\sum_\omega \mathsf{X}(\omega)\Pr(\omega)$, without affecting its value. In particular, equation (13) holds provided $\mathrm{E}[\mathsf{X}]$ exists, and equation (14) holds provided $\mathrm{E}[f(\mathsf{X})]$ exists.

Theorem 10 still holds, under the additional hypothesis that $\mathrm{E}[\mathsf{X}]$ and $\mathrm{E}[\mathsf{Y}]$ are finite. Equation (15) also holds, provided the individual expectations $\mathrm{E}[\mathsf{X}_i]$ are finite. More generally, if $\mathrm{E}[\mathsf{X}]$ and $\mathrm{E}[\mathsf{Y}]$ exist, then $\mathrm{E}[\mathsf{X}+\mathsf{Y}] = \mathrm{E}[\mathsf{X}] + \mathrm{E}[\mathsf{Y}]$, unless $\mathrm{E}[\mathsf{X}] = \infty$ and $\mathrm{E}[\mathsf{Y}] = -\infty$, or $\mathrm{E}[\mathsf{X}] = -\infty$ and $\mathrm{E}[\mathsf{Y}] = \infty$. Also, if $\mathrm{E}[\mathsf{X}]$ exists, then $\mathrm{E}[a\mathsf{X}] = a\,\mathrm{E}[\mathsf{X}]$, unless $a = 0$ and $\mathrm{E}[\mathsf{X}] = \pm\infty$.

One might consider generalizing (15) to countably infinite families of random variables. To this end, suppose $\{\mathsf{X}_i\}_{i=1}^\infty$ is an infinite sequence of real-valued random variables. The random variable $\mathsf{X} := \sum_{i=1}^\infty \mathsf{X}_i$ is well defined, provided the series $\sum_{i=1}^\infty \mathsf{X}_i(\omega)$ converges for each $\omega \in \Omega$. One might hope that $\mathrm{E}[\mathsf{X}] = \sum_{i=1}^\infty \mathrm{E}[\mathsf{X}_i]$; however, this is not in general true, even if the individual expectations, $\mathrm{E}[\mathsf{X}_i]$, are non-negative, and even if the series defining X converges absolutely for each ω; nevertheless, it is true when the X_i's are non-negative:

Theorem 22. *Let $\{\mathsf{X}_i\}_{i=1}^\infty$ be an infinite sequence of random variables. Suppose that for each $i \geq 1$, X_i takes non-negative values only, and has finite expectation. Also suppose that $\sum_{i=1}^\infty \mathsf{X}_i(\omega)$ converges for each $\omega \in \Omega$, and define $\mathsf{X} := \sum_{i=1}^\infty \mathsf{X}_i$. Then we have*

$$\mathrm{E}[\mathsf{X}] = \sum_{i=1}^\infty \mathrm{E}[\mathsf{X}_i].$$

Proof. This is a calculation just like the one made in the proof of Theorem 21, where, again, we use the fact that we may reverse the order of summation in an infinite double summation of non-negative terms:

$$\begin{aligned}\mathrm{E}[\mathsf{X}] &= \sum_{\omega\in\Omega} \Pr(\omega)\mathsf{X}(\omega) = \sum_{\omega\in\Omega} \Pr(\omega) \sum_{i=1}^\infty \mathsf{X}_i(\omega) \\ &= \sum_{i=1}^\infty \sum_{\omega\in\Omega} \Pr(\omega)\mathsf{X}_i(\omega) = \sum_{i=1}^\infty \mathrm{E}[\mathsf{X}_i]. \quad \square\end{aligned}$$

Theorem 11 holds under the additional hypothesis that $\mathrm{E}[\mathsf{X}]$ and $\mathrm{E}[\mathsf{Y}]$ are finite. Equation (16) also holds, provided the individual expectations $\mathrm{E}[\mathsf{X}_i]$ are finite. Theorem 12 still holds, of course. Theorem 13 also holds, but where now the sum may be infinite; the proof goes through essentially unchanged.

Example 44. Suppose X is a random variable with a geometric distribution, as in Example 40, with an associated success probability p and failure probability $q := 1-p$. As we saw in Example 42, for every integer $i \geq 1$, we have $\Pr[\mathsf{X} \geq i] = q^{i-1}$. We may therefore apply the infinite version of Theorem 13 to easily compute the expected value of X:

$$\mathrm{E}[\mathsf{X}] = \sum_{i=1}^\infty \Pr[\mathsf{X} \geq i] = \sum_{i=1}^\infty q^{i-1} = \frac{1}{1-q} = \frac{1}{p}. \quad \square$$

Example 45. To illustrate that Theorem 22 does not hold in general, consider the geometric distribution on the positive integers, where $\Pr(j) = 2^{-j}$ for $j \geq 1$. For $i \geq 1$, define the random variable X_i so that $\mathsf{X}_i(i) = 2^i$, $\mathsf{X}_i(i+1) = -2^{i+1}$, and $\mathsf{X}_i(j) = 0$ for all $j \notin \{i, i+1\}$. Then $\mathrm{E}[\mathsf{X}_i] = 0$ for all $i \geq 1$, and so $\sum_{i\geq 1} \mathrm{E}[\mathsf{X}_i] = 0$. Now define $\mathsf{X} := \sum_{i\geq 1} \mathsf{X}_i$. This is well defined, and in fact $\mathsf{X}(1) = 2$, while $\mathsf{X}(j) = 0$ for all $j > 1$. Hence $\mathrm{E}[\mathsf{X}] = 1$. $\square$

The definition of conditional expectation carries over verbatim. Equation (18) holds, provided $\mathrm{E}[\mathsf{X} \mid \mathcal{B}]$ exists, and the law of total expectation (19) holds, provided $\mathrm{E}[\mathsf{X}]$ exists. The law of total expectation also holds for a countably infinite partition $\{\mathcal{B}_i\}_{i \in I}$, provided $\mathrm{E}[\mathsf{X}]$ exists, and each of the conditional expectations $\mathrm{E}[\mathsf{X} \mid \mathcal{B}_i]$ is finite.

The variance $\mathrm{Var}[\mathsf{X}]$ of X exists only when $\mu := \mathrm{E}[\mathsf{X}]$ is finite, in which case it is defined as usual as $\mathrm{E}[(\mathsf{X} - \mu)^2]$, which may be either finite or infinite. Theorems 14, 15, and 16 hold provided all the relevant expectations and variances are finite.

6.5 Some useful bounds

All of the results in this section hold, provided the relevant expectations and variances are finite.

A Some useful facts

A1. Some handy inequalities. The following inequalities involving exponentials and logarithms are very handy.

(i) For all real numbers x, we have

$$1 + x \leq e^x,$$

or, taking logarithms, for $x > -1$, we have

$$\log(1 + x) \leq x.$$

(ii) For all real numbers $x \geq 0$, we have

$$e^{-x} \leq 1 - x + x^2/2,$$

or, taking logarithms,

$$-x \leq \log(1 - x + x^2/2).$$

Both (i) and (ii) follow easily from Taylor's formula with remainder, applied to the function e^x.

A2. Binomial coefficients. For integers n and k, with $0 \leq k \leq n$, one defines the **binomial coefficient**

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

We have the identities

$$\binom{n}{n} = \binom{n}{0} = 1,$$

and for $0 < k < n$, we have **Pascal's identity**

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k},$$

which may be verified by direct calculation. From these identities, it follows that $\binom{n}{k}$ is an integer, and indeed, is equal to the number of subsets of $\{1, \ldots, n\}$ of cardinality k. The usual

binomial theorem also follows as an immediate consequence: for all numbers a, b, and for all positive integers n, we have the **binomial expansion**

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

It is also easily verified, directly from the definition, that

$$\begin{aligned}
\binom{n}{k} &< \binom{n}{k+1} && \text{for } 0 \le k < (n-1)/2, \\
\binom{n}{k} &> \binom{n}{k+1} && \text{for } (n-1)/2 < k < n, \text{ and} \\
\binom{n}{k} &= \binom{n}{n-k} && \text{for } 0 \le k \le n.
\end{aligned}$$

In other words, if we fix n, and view $\binom{n}{k}$ as a function of k, then this function is increasing on the interval $[0, n/2]$, decreasing on the interval $[n/2, n]$, and its graph is symmetric with respect to the line $k = n/2$.

A3. Countably infinite sets. Let $\mathbb{Z}_{>0} := \{1, 2, 3, \ldots\}$, the set of positive integers. A set S is called **countably infinite** if there is a bijection $f : \mathbb{Z}_{>0} \to S$; in this case, we can enumerate the elements of S as $x_1, x_2, x_3, \ldots$, where $x_i := f(i)$.

A set S is called **countable** if it is either finite or countably infinite.

For a set S, the following conditions are equivalent:

- S is countable;
- there is a surjective function $g : \mathbb{Z}_{>0} \to S$;
- there is an injective function $h : S \to \mathbb{Z}_{>0}$.

The following facts can be easily established:

(i) if $S_1, \ldots, S_n$ are countable sets, then so are $S_1 \cup \cdots \cup S_n$ and $S_1 \times \cdots \times S_n$;

(ii) if $S_1, S_2, S_3, \ldots$ are countable sets, then so is $\bigcup_{i=1}^{\infty} S_i$;

(iii) if S is a countable set, then so is the set $\bigcup_{i=0}^{\infty} S^i$ of all finite sequences of elements in S.

Some examples of countably infinite sets: $\mathbb{Z}$, $\mathbb{Q}$, the set of all finite bit strings. Some examples of uncountable sets: $\mathbb{R}$, the set of all infinite bit strings.

A4. Infinite series. Consider an infinite series $\sum_{i=1}^{\infty} x_i$. It is a basic fact from calculus that if the x_i's are non-negative and $\sum_{i=1}^{\infty} x_i$ converges to a value y, then any infinite series whose terms are a rearrangement of the x_i's converges to the same value y.

If we drop the requirement that the x_i's are non-negative, but insist that the series $\sum_{i=1}^{\infty} |x_i|$ converges, then the series $\sum_{i=1}^{\infty} x_i$ is called **absolutely convergent**. In this case, then not only does the series $\sum_{i=1}^{\infty} x_i$ converge to some value y, but any infinite series whose terms are a rearrangement of the x_i's also converges to the same value y.

A5. Double infinite series. The topic of **double infinite series** may not be discussed in a typical introductory calculus course; we summarize here the basic facts that we need.

Suppose that $\{x_{ij}\}_{i,j=1}^{\infty}$ is a family non-negative real numbers such that for each i, the series $\sum_j x_{ij}$ converges to a value r_i, and for each j the series $\sum_i x_{ij}$ converges to a value c_j. Then we can form the double infinite series $\sum_i \sum_j x_{ij} = \sum_i r_i$ and the double infinite series $\sum_j \sum_i x_{ij} = \sum_j c_j$. If $(i_1, j_1), (i_2, j_2), \ldots$ is an enumeration of all pairs of indices (i, j), we can also form the single infinite series $\sum_k x_{i_k j_k}$. We then have $\sum_i \sum_j x_{ij} = \sum_j \sum_i x_{ij} = \sum_k x_{i_k j_k}$, where the three series either all converge to the same value, or all diverge. Thus, we can reverse the order of summation in a double infinite series of non-negative terms. If we drop the non-negativity requirement, the same result holds provided $\sum_k |x_{i_k j_k}| < \infty$.

Now suppose $\sum_i a_i$ is an infinite series of non-negative terms that converges to A, and that $\sum_j b_j$ is an infinite series of non-negative terms that converges to B. If $(i_1, j_1), (i_2, j_2), \ldots$ is an enumeration of all pairs of indices (i, j), then $\sum_k a_{i_k} b_{j_k}$ converges to AB. Thus, we can multiply term-wise infinite series with non-negative terms. If we drop the non-negativity requirement, the same result holds provided $\sum_i a_i$ and $\sum_j b_j$ converge absolutely.

A6. Convex functions. Let I be an interval of the real line (either open, closed, or half open, and either bounded or unbounded), and let f be a real-valued function defined on I. The function f is called **convex on** I if for all $x_0, x_2 \in I$, and for all $t \in [0, 1]$, we have

$$f(tx_0 + (1-t)x_2) \leq tf(x_0) + (1-t)f(x_2).$$

Geometrically, convexity means that for every three points $P_i = (x_i, f(x_i))$, $i = 0, 1, 2$, where each $x_i \in I$ and $x_0 < x_1 < x_2$, the point P_1 lies on or below the line through P_0 and P_2.

We state here the basic analytical facts concerning convex functions:

(i) if f is convex on I, then f is continuous on the interior of I (but not necessarily at the endpoints of I, if any);

(ii) if f is continuous on I and differentiable on the interior of I, then f is convex on I if and only if its derivative is non-decreasing on the interior of I.