A Primer on Probability

This is an abridged version of Chapter 8 of *A Computational Introduction to Number Theory and Algebra (2nd edition)*, by Victor Shoup.

1 Basic definitions

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

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

(1)

The set \( \Omega \) is called the sample space of \( P \).

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

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 \( P(\omega) := 1/6 \) for all \( \omega \in \Omega \), gives a probability distribution that naturally describes the possible outcomes of the experiment.

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

**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 \( A \) of \( \Omega \), and the probability of \( A \) is defined to be

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

(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 \( A \) and \( B \), their union \( A \cup B \) logically represents the event that *either* the event \( A \) or the event \( B \) occurs (or both), while their intersection \( A \cap B \) logically represents the event that *both* \( A \) and \( B \) occur. For an event \( A \), we define its complement \( \overline{A} := \Omega \setminus A \), which logically represents the event that \( A \) 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 \( A \) and \( B \),
\[
A \cup B = A \cap B \quad \text{and} \quad A \cap B = A \cup B.
\]
We also have the **Boolean distributive law**: for all events \( A \), \( B \), and \( C \),
\[
A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).
\]

**Example 4** Continuing with Example 1, the event that the die has an odd value is \( A := \{1, 3, 5\} \), and we have \( P[A] = 1/2 \). The event that the die has a value greater than 2 is \( B := \{3, 4, 5, 6\} \), and \( P[B] = 2/3 \). The event that the die has a value that is at most 2 is \( \overline{B} = \{1, 2\} \), and \( P[\overline{B}] = 1/3 \). The event that the value of the die is odd or exceeds 2 is \( A \cup B = \{1, 3, 4, 5, 6\} \), and \( P[A \cup B] = 5/6 \). The exact formula for arbitrary events \( A \) and \( B \) is:
\[
P[A \cup B] = P[A] + P[B] - P[A \cap B],
\]
and (3), (4), and (5) all follow from the observation that in the expression
\[
P[A] + P[B] = \sum_{\omega \in A} P(\omega) + \sum_{\omega \in B} P(\omega),
\]
the value \( P(\omega) \) is counted once for each \( \omega \in A \cup B \), except for those \( \omega \in A \cap B \), for which \( P(\omega) \) is counted twice.

**Example 5** If \( P \) is the uniform distribution on a set \( \Omega \), and \( A \) is a subset of \( \Omega \), then \( P[A] = |A|/|\Omega| \).

We next derive some elementary facts about probabilities of certain events, and relations among them. It is clear from the definitions that
\[
P[\emptyset] = 0 \quad \text{and} \quad P[\Omega] = 1,
\]
and that for every event \( A \), we have
\[
P[\overline{A}] = 1 - P[A].
\]

Now consider events \( A \) and \( B \), and their union \( A \cup B \). We have
\[
P[A \cup B] \leq P[A] + P[B],
\]
and that for every event \( A \), we have
\[
P[\overline{A}] = 1 - P[A].
\]

Moreover,
\[
P[A \cup B] = P[A] + P[B] \quad \text{if} \quad A \quad \text{and} \quad B \quad \text{are disjoint},
\]
that is, if \( A \cap B = \emptyset \). The exact formula for arbitrary events \( A \) and \( B \) is:
\[
P[A \cup B] = P[A] + P[B] - P[A \cap B],
\]
and (3), (4), and (5) all follow from the observation that in the expression
\[
P[A] + P[B] = \sum_{\omega \in A} P(\omega) + \sum_{\omega \in B} P(\omega),
\]
the value \( P(\omega) \) is counted once for each \( \omega \in A \cup B \), except for those \( \omega \in A \cap B \), for which \( P(\omega) \) 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 \( A_k \) be the event that the first die is \( k \), and \( B_k \) the event that the second die is \( k \). Let \( C_k = A_k \cup B_k \) be the event that \( k \) appears on either of the two dice. No matter what value \( k \) Bob chooses, the probability that this choice is correct is
\[
P[C_k] = P[A_k \cup B_k] = P[A_k] + P[B_k] - P[A_k \cap B_k]
\]
\[
= 1/6 + 1/6 - 1/36 = 11/36,
\]
which is slightly less than the estimate \( P[A_k] + P[B_k] \) obtained from (3).
Theorem 1 (Inclusion/exclusion principle) Let \(\{A_i\}_{i \in I}\) be a family of events, indexed by some set \(I\), we can naturally form the union \(\bigcup_{i \in I} A_i\) and intersection \(\bigcap_{i \in I} A_i\). If \(I = \emptyset\), then by definition, the union is \(\emptyset\), and by special convention, the intersection is the entire sample space \(\Omega\). Logically, the union represents the event that some \(A_i\) occurs, and the intersection represents the event that all the \(A_i\)'s occur. De Morgan's law generalizes as follows:
\[
\bigcup_{i \in I} A_i = \bigcap_{i \in I} A_i \quad \text{and} \quad \bigcap_{i \in I} A_i = \bigcup_{i \in I} A_i,
\]
and if \(B\) is an event, then the Boolean distributive law generalizes as follows:
\[
B \cap \left( \bigcup_{i \in I} A_i \right) = \bigcup_{i \in I} (B \cap A_i) \quad \text{and} \quad B \cup \left( \bigcap_{i \in I} A_i \right) = \bigcap_{i \in I} (B \cup A_i).
\]

We now generalize (3), (4), and (5) from pairs of events to families of events. Let \(\{A_i\}_{i \in I}\) be a finite family of events (i.e., the index set \(I\) is finite). Using (3), it follows by induction on \(|I|\) that
\[
P\left[ \bigcup_{i \in I} A_i \right] \leq \sum_{i \in I} P[A_i],
\]
which is known as **Boole's inequality** (and sometimes called the union bound). Analogously, using (4), it follows by induction on \(|I|\) that
\[
P\left[ \bigcup_{i \in I} A_i \right] = \sum_{i \in I} P[A_i] \quad \text{if} \quad \{A_i\}_{i \in I} \text{ is pairwise disjoint},
\]
that is, if \(A_i \cap A_j = \emptyset\) 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 \(A\) by breaking \(A\) up into a family \(\{A_i\}_{i \in I}\) of smaller, and hopefully simpler, events, whose union is \(A\). We shall make frequent use of them.

The generalization of (5) is messier. Consider first the case of three events, \(A, B,\) and \(C\). We have
\[
P[A \cup B \cup C] = P[A] + P[B] + P[C] - P[A \cap B] - P[A \cap C] - P[B \cap C] + P[A \cap B \cap C].
\]

Thus, starting with the sum of the probabilities of the individual events, we have to subtract a “correction term” that consists of the sum of probabilities of all intersections of pairs of events; however, this is an “over-correction,” and we have to correct the correction by adding back in the probability of the intersection of all three events. The general statement is as follows:

**Theorem 1 (Inclusion/exclusion principle)** Let \(\{A_i\}_{i \in I}\) be a finite family of events. Then
\[
P\left[ \bigcup_{i \in I} A_i \right] = \sum_{\emptyset \subseteq J \subseteq I} (-1)^{|J|-1} P\left[ \bigcap_{j \in J} A_j \right],
\]
the sum being over all non-empty subsets \(J\) of \(I\).

**Proof.** For \(\omega \in \Omega\) and \(B \subseteq \Omega\), define \(\delta_\omega [B] := 1\) if \(\omega \in B\), and \(\delta_\omega [B] := 0\) if \(\omega \notin B\). As a function of \(\omega\), \(\delta_\omega [B]\) is simply the characteristic function of \(B\). One may easily verify that for all \(\omega \in \Omega\), \(B \subseteq \Omega\), and \(C \subseteq \Omega\), we have \(\delta_\omega [B] = 1 \delta_\omega [B] \) and \(\delta_\omega [B \cap C] = \delta_\omega [B] \delta_\omega [C]\). It is also easily seen that for every \(B \subseteq \Omega\), we have \(\sum_{\omega \in \Omega} P(\omega) \delta_\omega [B] = P[B]\).
Let $A := \bigcup_{i \in I} A_i$, and for $J \subset I$, let $A_J := \bigcap_{j \in J} A_j$. For every $\omega \in \Omega$,

$$1 - \delta_\omega[A] = \delta_\omega[\mathcal{A}] = \delta_\omega\bigcap_{i \in I} A_i = \prod_{i \in I} \delta_\omega[A_i] = \prod_{i \in I} (1 - \delta_\omega[A_i])$$

$$= \sum_{J \subset I} (-1)^{|J|} \prod_{j \in J} \delta_\omega[A_j] = \sum_{J \subset I} (-1)^{|J|} \delta_\omega[A_J],$$

and so

$$\delta_\omega[A] = \sum_{\emptyset \subseteq J \subset I} (-1)^{|J| - 1} \delta_\omega[A_J]. \quad (8)$$

Multiplying (8) by $P(\omega)$, and summing over all $\omega \in \Omega$, we have

$$P[A] = \sum_{\omega \in \Omega} P(\omega) \delta_\omega[A] = \sum_{\omega \in \Omega} P(\omega) \sum_{\emptyset \subseteq J \subset I} (-1)^{|J| - 1} \delta_\omega[A_J]$$

$$= \sum_{\emptyset \subseteq J \subset I} (-1)^{|J|-1} \sum_{\omega \in \Omega} P(\omega) \delta_\omega[A_J] = \sum_{\emptyset \subseteq J \subset I} (-1)^{|J|-1} P[A_J].$$

One can also state the inclusion/exclusion principle in a slightly different way, splitting the sum into terms with $|J| = 1$, $|J| = 2$, etc., as follows:

$$P\left[\bigcup_{i \in I} A_i\right] = \sum_{i \in I} P[A_i] + \sum_{k=2}^{|I|} (-1)^{k-1} \sum_{J \subset I, |J|=k} P\left[\bigcap_{j \in J} A_j\right],$$

where the last sum in this formula is taken over all subsets $J$ of $I$ of size $k$.

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 in general, as follows.

Let $P_1 : \Omega_1 \to [0,1]$ and $P_2 : \Omega_2 \to [0,1]$ be probability distributions. Their **product distribution** $P := P_1 P_2$ is defined as follows:

$$P : \Omega_1 \times \Omega_2 \to [0,1]$$

$$(\omega_1, \omega_2) \mapsto P_1(\omega_1) P_2(\omega_2).$$

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

$$\sum_{\omega_1, \omega_2} P(\omega_1, \omega_2) = \sum_{\omega_1, \omega_2} P_1(\omega_1) P_2(\omega_2) = \left(\sum_{\omega_1} P_1(\omega_1)\right) \left(\sum_{\omega_2} P_2(\omega_2)\right) = 1 \cdot 1 = 1.$$

More generally, if $P_i : \Omega_i \to [0,1]$, for $i = 1, \ldots, n$, are probability distributions, then their product distribution is $P := P_1 \cdots P_n$, where

$$P : \Omega_1 \times \cdots \times \Omega_n \to [0,1]$$

$$(\omega_1, \ldots, \omega_n) \mapsto P_1(\omega_1) \cdots P_n(\omega_n).$$

If $P_1 = P_2 = \cdots = P_n$, then we may write $P = P_1^n$. It is clear from the definitions that if each $P_i$ is the uniform distribution on $\Omega_i$, then $P$ is the uniform distribution on $\Omega_1 \times \cdots \times \Omega_n$. 

4
Example 7 We can view the probability distribution $P$ in Example 6 as $P_1^2$, where $P_1$ is the uniform distribution on $\{1, \ldots, 6\}$.

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 $P = P_1^n$, where $P_1$ 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 $\Omega$ of $P$ is the set of all $2^n$ tuples $\omega = (\omega_1, \ldots, \omega_n)$, where each $\omega_i$ is either heads or tails. If the tuple $\omega$ has $k$ heads and $n - k$ tails, then $P(\omega) = p^k q^{n-k}$, regardless of the positions of the heads and tails in the tuple.

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

$$|A_k| = \binom{n}{k},$$

from which it follows that

$$P[A_k] = \binom{n}{k} p^k q^{n-k}.$$

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

$$P[A_k] = \binom{n}{k} 2^{-n}.$$

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

Suppose $P : \Omega \to [0, 1]$ is a probability distribution. The support of $P$ is defined to be the set $\{\omega \in \Omega : P(\omega) \neq 0\}$. Now consider another probability distribution $P' : \Omega' \to [0, 1]$. Of course, these two distributions are equal if and only if $\Omega = \Omega'$ and $P(\omega) = P'(\omega)$ for all $\omega \in \Omega$. However, it is natural and convenient to have a more relaxed notion of equality. We shall say that $P$ and $P'$ are essentially equal if the restriction of $P$ to its support is equal to the restriction of $P'$ to its support. For example, if $P$ 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 $P$ is essentially the uniform distribution on $\{1, 2, 3\}$.

Exercise 1 Show that $P[A \cap B] P[A \cup B] \leq P[A] P[B]$ for all events $A, B$.

Exercise 2 Suppose $A, B, C$ are events such that $A \cap \overline{C} = B \cap \overline{C}$. Show that $|P[A] - P[B]| \leq P[C]$.

Exercise 3 This exercise asks you to generalize Boole’s inequality (6), proving Bonferroni’s inequalities. Let $\{A_i\}_{i \in I}$ be a finite family of events, where $n := |I|$. For $m = 0, \ldots, n$, define

$$\alpha_m := \sum_{k=1}^{m} (-1)^{k-1} \sum_{|J|=k} \left[ \prod_{j \in J} P[A_j] \right].$$

Also, define

$$\alpha := \prod_{i \in I} P[A_i].$$

Show that $\alpha \leq \alpha_m$ if $m$ is odd, and $\alpha \geq \alpha_m$ if $m$ is even. Hint: use induction on $n$.  

5
2 Conditional probability and independence

Let \( \mathbb{P} \) be a probability distribution on a sample space \( \Omega \).

For a given event \( \mathcal{B} \subset \Omega \) with \( \mathbb{P}[\mathcal{B}] \neq 0 \), and for \( \omega \in \Omega \), let us define
\[
\mathbb{P}(\omega \mid \mathcal{B}) := \begin{cases} \frac{\mathbb{P}(\omega)}{\mathbb{P}[\mathcal{B}]} & \text{if } \omega \in \mathcal{B}, \\ 0 & \text{otherwise}. \end{cases}
\]

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

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

For a given event \( \mathcal{A} \subset \Omega \), its probability with respect to the conditional distribution given \( \mathcal{B} \) is
\[
\mathbb{P}[(\mathcal{A} \cap \mathcal{B})] = \frac{\mathbb{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]}.
\]
The value \( \mathbb{P}[(\mathcal{A} \mid \mathcal{B})] \) 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 \( \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}] \mathbb{P}[\mathcal{B}] \), then \( \mathcal{A} \) and \( \mathcal{B} \) are called independent events. If \( \mathbb{P}[\mathcal{B}] \neq 0 \), one easily sees that \( \mathcal{A} \) and \( \mathcal{B} \) are independent if and only if \( \mathbb{P}[(\mathcal{A} \mid \mathcal{B})] = \mathbb{P}[(\mathcal{A})] \); 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 \( \mathbb{P} \) is the uniform distribution on \( \Omega \), and that \( \mathcal{B} \subset \Omega \) with \( \mathbb{P}[\mathcal{B}] \neq 0 \). 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, \( \mathbb{P}[\mathcal{A}] = 1/2 \), \( \mathbb{P}[\mathcal{B}] = 2/3 \), and \( \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = 1/3 \); thus, \( \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}] \mathbb{P}[\mathcal{B}] \), and we conclude that \( \mathcal{A} \) and \( \mathcal{B} \) are independent. Indeed, \( \mathbb{P}[(\mathcal{A} \mid \mathcal{B})] = (1/3)/(2/3) = 1/2 = \mathbb{P}[(\mathcal{A})] \); 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 \( \mathbb{P}[\mathcal{C}] = 1/2 \) and \( \mathbb{P}[\mathcal{A} \cap \mathcal{C}] = 1/6 \neq 1/4 \), from which we conclude that \( \mathcal{A} \) and \( \mathcal{C} \) are not independent. Indeed, \( \mathbb{P}[(\mathcal{A} \mid \mathcal{C})] = (1/6)/(1/2) = 1/3 \neq \mathbb{P}[(\mathcal{A})] \); 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:

\[
\begin{array}{cccccccccccc}
6 & | & 7 & 8 & 9 & 10 & 11 & 12 \\
5 & | & 6 & 7 & 8 & 9 & 10 & \\
4 & | & 5 & 6 & 7 & 8 & 9 & \\
3 & | & 4 & 5 & 6 & 7 & 8 & \\
2 & | & 3 & 4 & 5 & 6 & 7 & \\
1 & | & 2 & 3 & 4 & 5 & 6 & \\
\end{array}
\]

\[
1 \ 2 \ 3 \ 4 \ 5 \ 6
\]

6
For example, suppose Alice tells Bob the sum is 4. Then what is Bob’s best strategy in this case? Let $D_\ell$ be the event that the sum is $\ell$, for $\ell = 2, \ldots, 12$, and consider the conditional distribution given $D_4$. 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, 

$$P[C_1 \mid D_4] = P[C_3 \mid D_4] = \frac{2}{3},$$

while

$$P[C_2 \mid D_4] = \frac{1}{3}$$

and

$$P[C_4 \mid D_4] = P[C_5 \mid D_4] = P[C_6 \mid D_4] = 0.$$ 

Thus, if the sum is 4, Bob’s best strategy is to guess either 1 or 3, which will be correct with probability $2/3$.

Similarly, if the sum is 5, then we consider the conditional distribution given $D_5$, 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 $P[C_k \mid D_5] = \frac{1}{2}$.

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

$$P[A] = \sum_{i \in I} P[A \cap B_i].$$

(9)

Furthermore, if each $B_i$ occurs with non-zero probability (so that, in particular, $\{B_i\}_{i \in I}$ is a partition of $\Omega$), then we have

$$P[A] = \sum_{i \in I} P[A \mid B_i] P[B_i].$$

(10)

If, in addition, $P[A] \neq 0$, then for each $j \in I$, we have

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] P[B_j]}{\sum_{i \in I} P[A \mid B_i] P[B_i]}.$$ 

(11)

Equations (9) and (10) are sometimes called the law of total probability, while equation (11) is known as Bayes’ theorem. Equation (10) (resp., (11)) is useful for computing or estimating $P[A]$ (resp., $P[B_j \mid A]$) by conditioning on the events $B_i$.

**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 $\ell$, and there are $N_\ell$ pairs in the sample space that sum to that number, then there will always be a value that appears in exactly 2 of these $N_\ell$ 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 $C$ is the event that Bob wins following this strategy, then by total probability (10), we have

$$P[C] = \sum_{\ell=2}^{12} P[C \mid D_\ell] P[D_\ell].$$
Moreover, 
\[ P[C \mid D_2] P[D_2] = 1 \cdot \frac{1}{36} = \frac{1}{36}, \quad P[C \mid D_{12}] P[D_{12}] = 1 \cdot \frac{1}{36} = \frac{1}{36}, \]
and for \( \ell = 3, \ldots, 11 \), we have 
\[ P[C \mid D_\ell] P[D_\ell] = \frac{2}{N_\ell} \cdot \frac{N_\ell}{36} = \frac{1}{18}. \]
Therefore, 
\[ P[C] = \frac{1}{36} + \frac{1}{36} + \frac{9}{18} = \frac{10}{18}. \]

**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: it has a 5% false positive rate (i.e., 5% of healthy patients test positive for the disease), and a 2% false negative rate (i.e., 2% of sick patients test negative for the disease). 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 \( A \) be the event that the test is positive and let \( B \) be the event that the patient has disease \( X \). The relevant quantity that we need to estimate is \( P[B \mid A] \); that is, the probability that the patient has disease \( X \), given a positive test result. We use Bayes’ theorem to do this:

\[ P[B \mid A] = \frac{P[A \mid B] P[B]}{P[A \mid B] P[B] + P[A \mid \overline{B}] P[\overline{B}]} = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.17. \]

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 1970’s. 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, Monty Hall, 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 \( B \) be the event that the contestant’s initial choice was correct, and let \( A \) 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
In this case, Monty has two doors to choose from, but his choice does not affect the outcome. Thus, \( P[A \mid B] = 0 \). 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 Monty reveals the other zonk, switching will lead with certainty to success. Thus, \( P[A \mid \overline{B}] = 1 \). Furthermore, it is clear that \( P[B] = 1/3 \). So using total probability (10), we compute

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

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. Monty did not always reveal a zonk and offer a choice to switch. Indeed, if Monty only revealed a zonk when the contestant had chosen the correct door, then switching would certainly be the wrong strategy. However, if Monty’s choice itself was a random decision made independently of the contestant’s initial choice, then switching is again the preferred strategy.

We next generalize the notion of independence from pairs of events to families of events. Let \( \{A_j\}_{i \in I} \) be a finite family of events. For a given positive integer \( k \), we say that the family \( \{A_i\}_{i \in I} \) is \( k \)-wise independent if the following holds:

\[
P\left[ \bigcap_{j \in J} A_j \right] = \prod_{j \in J} P[A_j] \quad \text{for all } J \subset I \text{ with } |J| \leq k.
\]

The family \( \{A_i\}_{i \in I} \) is called pairwise independent if it is 2-wise independent. Equivalently, pairwise independence means that for all \( i, j \in I \) with \( i \neq j \), we have \( P[A_i \cap A_j] = P[A_i] P[A_j] \), or put yet another way, that for all \( i, j \in I \) with \( i \neq j \), the events \( A_i \) and \( A_j \) are independent.

The family \( \{A_i\}_{i \in I} \) is called mutually independent if it is \( k \)-wise independent for all positive integers \( k \). Equivalently, mutual independence means that

\[
P\left[ \bigcap_{j \in J} A_j \right] = \prod_{j \in J} P[A_j] \quad \text{for all } J \subset I.
\]

If \( n := |I| > 0 \), mutual independence is equivalent to \( n \)-wise independence; moreover, if \( 0 < k \leq n \), then \( \{A_i\}_{i \in I} \) is \( k \)-wise independent if and only if \( \{A_j\}_{j \in J} \) is mutually independent for every \( J \subset I \) with \( |J| = k \).

In defining independence, the choice of the index set \( I \) plays no real role, and we can rename elements of \( I \) as convenient.

**Example 15** Suppose we toss a fair coin three times, which we formally model using the uniform distribution on the set of all 8 possible outcomes of the three coin tosses: (heads, heads, heads), (heads, heads, tails), etc., as in Example 8. For \( i = 1, 2, 3 \), let \( A_i \) be the event that the \( i \)th toss comes up heads. Then \( \{A_i\}_{i=1}^3 \) is a mutually independent family of events, where each individual \( A_i \) occurs with probability 1/2.

Now let \( B_{12} \) be the event that the first and second tosses agree (i.e., both heads or both tails), let \( B_{13} \) be the event that the first and third tosses agree, and let \( B_{23} \) be the event that the second and third tosses agree. Then the family of events \( B_{12}, B_{13}, B_{23} \) is pairwise independent, but not mutually independent. Indeed, the probability that any given individual event occurs is 1/2, and the probability that any given pair of events occurs is 1/4; however, the probability that all three events occur is also 1/4, since if any two events occur, then so does the third.
We close this section with some simple facts about independence of events and their complements.

**Theorem 2** If $A$ and $B$ are independent events, then so are $A$ and $\overline{B}$.

**Proof.** We have

$$P[A] = P[A \cap B] + P[A \cap \overline{B}] \quad \text{(by total probability (9))}$$

$$= P[A]P[B] + P[A \cap \overline{B}] \quad \text{(since $A$ and $B$ are independent).}$$

Therefore,

$$P[A \cap \overline{B}] = P[A] - P[A]P[B] = P[A](1 - P[B]) = P[A]P[\overline{B}].$$

□

This theorem implies that

$$A \text{ and } B \text{ are independent } \iff \overline{A} \text{ and } B \text{ " } \iff \overline{A} \text{ and } \overline{B} \text{ " } \iff A \text{ and } \overline{B} \text{ " }.$$

The following theorem generalizes this result to families of events. It says that if a family of events is $k$-wise independent, then the family obtained by complementing any number of members of the given family is also $k$-wise independent.

**Theorem 3** Let $\{A_i\}_{i \in I}$ be a finite, $k$-wise independent family of events. Let $J$ be a subset of $I$, and for each $i \in I$, define $A'_i := A_i$ if $i \in J$, and $A'_i := \overline{A}_i$ if $i \notin J$. Then $\{A'_i\}_{i \in I}$ is also $k$-wise independent.

**Proof.** It suffices to prove the theorem for the case where $J = I \setminus \{d\}$, for an arbitrary $d \in I$: this allows us to complement any single member of the family that we wish, without affecting independence; by repeating the procedure, we can complement any number of them.

To this end, it will suffice to show the following: if $J \subset I$, $|J| < k$, $d \in I \setminus J$, and $A_J := \bigcap_{j \in J} A_j$, we have

$$P[\overline{A}_d \cap A_J] = (1 - P[A_d]) \prod_{j \in J} P[A_j]. \quad (12)$$

Using total probability (9), along with the independence hypothesis (twice), we have

$$\prod_{j \in J} P[A_j] = P[A_J] = P[A_d \cap A_J] + P[\overline{A}_d \cap A_J]$$

$$= P[A_d] \cdot \prod_{j \in J} P[A_j] + P[\overline{A}_d \cap A_J],$$

from which (12) follows immediately. □

**Exercise 4** For events $A_1, \ldots, A_n$, define $\alpha_1 := P[A_1]$, and for $i = 2, \ldots, n$, define $\alpha_i := P[A_i \mid A_1 \cap \cdots \cap A_{i-1}]$ (assume that $P[A_1 \cap \cdots \cap A_{n-1}] \neq 0$). Show that $P[A_1 \cap \cdots \cap A_n] = \alpha_1 \cdots \alpha_n$.  

10
Exercise 5 Let $\mathcal{B}$ be an event, and let $\{\mathcal{B}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union is $\mathcal{B}$. Generalizing the law of total probability (equations (9) and (10)), show that for every event $\mathcal{A}$, we have $P[\mathcal{A} \cap \mathcal{B}] = \sum_{i \in I} P[\mathcal{A} \cap \mathcal{B}_i]$, and if $P[\mathcal{B}] \neq 0$ and $I^* := \{i \in I : P[\mathcal{B}_i] \neq 0\}$, then

$$P[\mathcal{A} | \mathcal{B}] P[\mathcal{B}] = \sum_{i \in I^*} P[\mathcal{A} | \mathcal{B}_i] P[\mathcal{B}_i].$$

Also show that if $P[\mathcal{A} | \mathcal{B}_i] \leq \alpha$ for each $i \in I^*$, then $P[\mathcal{A} | \mathcal{B}] \leq \alpha$.

Exercise 6 Let $\mathcal{B}$ be an event with $P[\mathcal{B}] \neq 0$, and let $\{\mathcal{C}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union contains $\mathcal{B}$. Again, generalizing the law of total probability, show that for every event $\mathcal{A}$, if $I^* := \{i \in I : P[\mathcal{B} \cap \mathcal{C}_i] \neq 0\}$, then we have

$$P[\mathcal{A} | \mathcal{B}] = \sum_{i \in I^*} P[\mathcal{A} | \mathcal{B} \cap \mathcal{C}_i] P[\mathcal{C}_i | \mathcal{B}].$$

Exercise 7 Three fair coins are tossed. Let $A$ be the event that at least two coins are heads. Let $B$ be the event that the number of heads is odd. Let $C$ be the event that the third coin is heads. Are $A$ and $B$ independent? $A$ and $C$? $B$ and $C$?

Exercise 8 Consider again the situation in Example 11, but now suppose that Alice only tells Bob the value of the sum of the two dice modulo 6. Describe an optimal strategy for Bob, and calculate his overall probability of winning.

Exercise 9 Consider again the situation in Example 13, but now suppose that the patient is visiting the doctor because he has symptom $Y$. Furthermore, it is known that everyone who has disease $X$ exhibits symptom $Y$, while 10\% of the population overall exhibits symptom $Y$. Assuming that the accuracy of the test is not affected by the presence of symptom $Y$, how should the doctor advise his patient should the test come out positive?

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 $P$ be a probability distribution on a sample space $\Omega$. A random variable $X$ is a function $X : \Omega \to S$, 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 $s \in S$, “$X = s$” denotes the event $\{\omega \in \Omega : X(\omega) = s\}$. It is immediate from this definition that

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

More generally, for any predicate $\phi$ on $S$, we may write “$\phi(X)$” as shorthand for the event $\{\omega \in \Omega : \phi(X(\omega))\}$. When we speak of the image of $X$, we simply mean its image in the usual function-theoretic sense, that is, the set $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. 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 $X_1, \ldots, X_n$ are random variables, where $X_i : \Omega \to S_i$ for $i = 1, \ldots, n$. Then $(X_1, \ldots, X_n)$ denotes the random
variable that maps $\omega \in \Omega$ to $(X_1(\omega), \ldots, X_n(\omega)) \in S_1 \times \cdots \times S_n$. If $f : S_1 \times \cdots \times S_n \to T$ is a function, then $f(X_1, \ldots, X_n)$ denotes the random variable that maps $\omega \in \Omega$ to $f(X_1(\omega), \ldots, X_n(\omega))$. 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 $X \star Y$ denotes the random variable that maps $\omega \in \Omega$ to $X(\omega) \star Y(\omega) \in S$.

Let $X$ be a random variable whose image is $S$. The variable $X$ determines a probability distribution $P_X : S \to [0, 1]$ on the set $S$, where $P_X(s) := P[X = s]$ for each $s \in S$. We call $P_X$ the distribution of $X$. If $P_X$ 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 $P[X = s] = P[Y = s]$ for all $s \in S$, then the distributions of $X$ and $Y$ are essentially equal even if their images are not identical.

Example 16 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 $\Omega$, where

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

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

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

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

<table>
<thead>
<tr>
<th>$u$</th>
<th>$2$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
<th>$6$</th>
<th>$7$</th>
<th>$8$</th>
<th>$9$</th>
<th>$10$</th>
<th>$11$</th>
<th>$12$</th>
</tr>
</thead>
</table>

Example 17 If $A$ is an event, we may define a random variable $X$ as follows: $X := 1$ if the event $A$ occurs, and $X := 0$ otherwise. The variable $X$ is called the indicator variable for $A$. Formally, $X$ is the function that maps $\omega \in A$ to 1, and $\omega \in \Omega \setminus A$ to 0; that is, $X$ is simply the characteristic function of $A$. The distribution of $X$ is that of a Bernoulli trial: $P[X = 1] = P[A]$ and $P[X = 0] = 1 - P[A]$.

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

Example 18 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 $A_i$ be the event that the $i$th toss comes up heads, and let $X_i$ be the corresponding indicator variable. Let us also define $X := X_1 + \cdots + X_n$, 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

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

12
The distribution of the random variable $X$ is called a \textbf{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 \textbf{regular function} if every element in the image of $f$ has the same number of pre-images under $f$.

**Theorem 4** Suppose $f : S \to T$ is a surjective, regular function, and that $X$ is a random variable that is uniformly distributed over $S$. Then $f(X)$ 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

\[
P[f(X) = t] = \sum_{\omega \in X^{-1}(S_t)} P(\omega) = \sum_{s \in S_t} \sum_{\omega \in X^{-1}(\{s\})} P(\omega) = \sum_{s \in S_t} P[X = s]
\]

\[
= \sum_{s \in S_t} 1/|S| = (|S|/|T|)/|S| = 1/|T|.
\]

$\Box$

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

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

\[
P[(X = s) \cap (Y = t)] = P[X = s] P[Y = t].
\]

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

\[
P[X = s \mid Y = t] = P[X = s]
\]

for all $s \in S$; that is, the conditional distribution of $X$ given $Y = t$ 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 19** Let us continue with Example 16. 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 $Z = 4$. We have $P[X = s \mid Z = 4] = 1/3$ for $s = 1, 2, 3$, and $P[X = s \mid Z = 4] = 0$ for $s = 4, 5, 6$. Thus, the conditional distribution of $X$ given $Z = 4$ is essentially the uniform distribution on $\{1, 2, 3\}$. Let us calculate
the conditional distribution of $Z$ given $X = 1$. We have $P[Z = u \mid X = 1] = 1/6$ for $u = 2, \ldots, 7$, and $P[Z = u \mid X = 1] = 0$ for $u = 8, \ldots, 12$. Thus, the conditional distribution of $Z$ given $X = 1$ is essentially the uniform distribution on $\{2, \ldots, 7\}$. In particular, it is clear that $X$ and $Z$ are not independent.

But now consider the random variable $Z' = (X+Y) \mod 6$. Fix any $s = 1, \ldots, 6$ and $u = 0, \ldots, 5$. We have

$$P[(Z' = u) \cap (X = s)] = P[(X + Y \equiv u \mod 6) \cap (X = s)] = P[(s + Y \equiv u \mod 6) \cap (X = s)]$$

$$= P[(Y \equiv u - s \mod 6) \cap (X = s)] = P[(Y = t) \cap (X = s)],$$

where $t$ is the unique integer in the interval $1, \ldots, 6$ such that $t \equiv u - s \mod 6$. Since $X$ and $Y$ are independent, it follows that

$$P[(Z' = u) \cap (X = s)] = P[(Y = t) \cap (X = s)] = P[Y = t]P[X = s] = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$  

Since this holds for all $s = 1, \ldots, 6$, we may use the law of total probability to calculate

$$P[Z' = u] = \sum_{s=1}^{6} P[(Z' = u) \cap (X = s)] = \sum_{s=1}^{6} \frac{1}{36} = \frac{1}{6}.$$  

It follows that $Z'$ is uniformly distributed over $\{0, \ldots, 5\}$, and that $Z'$ and $X$ are independent.

We now generalize the notion of independence to families of random variables. Let $\{X_i\}_{i \in I}$ be a finite family of random variables. Let us call a corresponding family of values $\{s_i\}_{i \in I}$ an assignment to $\{X_i\}_{i \in I}$ if $s_i$ is in the image of $X_i$ for each $i \in I$. For a given positive integer $k$, we say that the family $\{X_i\}_{i \in I}$ is $k$-wise independent if for every assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$, the family of events $\{X_i = s_i\}_{i \in I}$ is $k$-wise independent.

The notions of pairwise and mutual independence for random variables are defined following the same pattern that was used for events. The family $\{X_i\}_{i \in I}$ is called pairwise independent if it is 2-wise independent, which means that for all $i, j \in I$ with $i \neq j$, the variables $X_i$ and $X_j$ are independent. The family $\{X_i\}_{i \in I}$ is called mutually independent if it is $k$-wise independent for all positive integers $k$. Equivalently, and more explicitly, mutual independence means that for every assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$, we have

$$P \left[ \bigcap_{j \in J} (X_j = s_j) \right] = \prod_{j \in J} P[X_j = s_j] \text{ for all } J \subseteq I. \quad (13)$$

If $n := |I| > 0$, mutual independence is equivalent to $n$-wise independence; moreover, if $0 < k \leq n$, then $\{X_i\}_{i \in I}$ is $k$-wise independent if and only if $\{X_j\}_{j \in J}$ is mutually independent for every $J \subseteq I$ with $|J| = k$.

**Example 20** Suppose $\{A_i\}_{i \in I}$ is a finite family of events. Let $\{X_i\}_{i \in I}$ be the corresponding family of indicator variables, so that for each $i \in I$, $X_i = 1$ if $A_i$ occurs, and $X_i = 0$, otherwise. Then immediately implies that for every positive integer $k$, $\{A_i\}_{i \in I}$ is $k$-wise independent if and only if $\{X_i\}_{i \in I}$ is $k$-wise independent.

**Example 21** Consider again Example 15, where we toss a fair coin 3 times. For $i = 1, 2, 3$, let $X_i$ be the indicator variable for the event $A_i$ that the $i$th toss comes up heads. Then $\{X_i\}_{i=1}^3$ is a mutually independent family of random variables. Let $Y_{12}$ be the indicator variable for the event $B_{12}$ that tosses 1 and 2 agree; similarly, let $Y_{13}$ be the indicator variable for the event $B_{13}$, and $Y_{23}$ the indicator variable for $B_{23}$. Then the family of random variables $Y_{12}, Y_{13}, Y_{23}$ is pairwise independent, but not mutually independent.
We next present a number of useful tools for establishing independence.

**Theorem 5** Let $X$ be a random variable with image $S$, and $Y$ be a random variable with image $T$. Further, suppose that $f : S \to [0,1]$ and $g : T \to [0,1]$ are functions such that

$$\sum_{s \in S} f(s) = \sum_{t \in T} g(t) = 1, \quad (14)$$

and that for all $s \in S$ and $t \in T$, we have

$$P[(X = s) \cap (Y = t)] = f(s)g(t). \quad (15)$$

Then $X$ and $Y$ are independent, the distribution of $X$ is $f$, and the distribution of $Y$ is $g$.

**Proof.** Since $\{Y = t\}_{t \in T}$ is a partition of the sample space, making use of total probability (9), along with (15) and (14), we see that for all $s \in S$, we have

$$P[X = s] = \sum_{t \in T} P[(X = s) \cap (Y = t)] = \sum_{t \in T} f(s)g(t) = f(s)\sum_{t \in T} g(t) = f(s).$$

Thus, the distribution of $X$ is indeed $f$. Exchanging the roles of $X$ and $Y$ in the above argument, we see that the distribution of $Y$ is $g$. Combining this with (15), we see that $X$ and $Y$ are independent. $\square$

The generalization of Theorem 5 to families of random variables is a bit messy, but the basic idea is the same:

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

$$P[\bigcap_{i \in I} (X_i = s_i)] = \prod_{i \in I} f_i(s_i)$$

for each assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$. Then the family $\{X_i\}_{i \in I}$ is mutually independent, and for each $i \in I$, the distribution of $X_i$ is $f_i$.

**Proof.** To prove the theorem, it suffices to prove the following statement: for every subset $J$ of $I$, and every assignment $\{s_j\}_{j \in J}$ to $\{X_j\}_{j \in J}$, we have

$$P[\bigcap_{j \in J} (X_j = s_j)] = \prod_{j \in J} f_j(s_j).$$

Moreover, it suffices to prove this statement for the case where $J = I \setminus \{d\}$, for an arbitrary $d \in I$: this allows us to eliminate any one variable from the family, without affecting the hypotheses, and by repeating this procedure, we can eliminate any number of variables.

Thus, let $d \in I$ be fixed, let $J := I \setminus \{d\}$, and let $\{s_j\}_{j \in J}$ be a fixed assignment to $\{X_j\}_{j \in J}$. Then, since $\{X_d = s_d\}_{s_d \in S_d}$ is a partition of the sample space, we have

$$P[\bigcap_{j \in J} (X_j = s_j)] = P[\bigcup_{s_d \in S_d} \left(\bigcap_{i \in I} (X_i = s_i)\right)] = \sum_{s_d \in S_d} P[\bigcap_{i \in I} (X_i = s_i)]$$

$$= \sum_{s_d \in S_d} \prod_{i \in I} f_i(s_i) = \prod_{j \in J} f_j(s_j) \cdot \sum_{s_d \in S_d} f_d(s_d) = \prod_{j \in J} f_j(s_j).$$

$\square$
This theorem has several immediate consequences. First of all, mutual independence may be more simply characterized:

**Theorem 7** Let \( \{X_i\}_{i \in I} \) be a finite family of random variables. Suppose that for every assignment \( \{s_i\}_{i \in I} \) to \( \{X_i\}_{i \in I} \), we have

\[
P\left[ \bigcap_{i \in I} (X_i = s_i) \right] = \prod_{i \in I} P[X_i = s_i].
\]

Then \( \{X_i\}_{i \in I} \) is mutually independent.

Theorem 7 says that to check for mutual independence, we only have to consider the index set \( J = I \) in (13). Put another way, it says that a family of random variables \( \{X_i\}_{i = 1}^n \) is mutually independent if and only if the distribution of \( (X_1, \ldots, X_n) \) is essentially equal to the product of the distributions of the individual \( X_i \)'s.

Based on the definition of mutual independence, and its characterization in Theorem 7, the following is also immediate:

**Theorem 8** Suppose \( \{X_i\}_{i = 1}^n \) is a family of random variables, and that \( m \) is an integer with \( 0 < m < n \). Then the following are equivalent:

(i) \( \{X_i\}_{i = 1}^n \) is mutually independent;

(ii) \( \{X_i\}_{i = 1}^m \) is mutually independent, \( \{X_i\}_{i = m+1}^n \) is mutually independent, and the two variables \( (X_1, \ldots, X_m) \) and \( (X_{m+1}, \ldots, X_n) \) are independent.

The following is also an immediate consequence of Theorem 6 (it also follows easily from Theorem 4).

**Theorem 9** Suppose that \( X_1, \ldots, X_n \) are random variables, and that \( S_1, \ldots, S_n \) are finite sets. Then the following are equivalent:

(i) \( (X_1, \ldots, X_n) \) is uniformly distributed over \( S_1 \times \cdots \times S_n \);

(ii) \( \{X_i\}_{i = 1}^n \) is mutually independent, with each \( X_i \) uniformly distributed over \( S_i \).

Another immediate consequence of Theorem 6 is the following:

**Theorem 10** Suppose \( P \) is the product distribution \( P_1 \cdots P_n \), where each \( P_i \) is a probability distribution on a sample space \( \Omega_i \), so that the sample space of \( P \) 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 \( X_i(\omega_1, \ldots, \omega_n) = \omega_i \). Then \( \{X_i\}_{i = 1}^n \) is mutually independent, and for each \( i = 1, \ldots, n \), the distribution of \( X_i \) is \( P_i \).

Theorem 10 is often used to synthesize independent random variables “out of thin air,” by taking the product of appropriate probability distributions. Other arguments may then be used to prove the independence of variables derived from these.

**Example 22** Theorem 10 immediately implies that in Example 18, the family of indicator variables \( \{X_i\}_{i = 1}^n \) is mutually independent.

The following theorem gives us yet another way to establish independence.
Theorem 11 Suppose $\{X_i\}_{i=1}^n$ is a mutually independent family of random variables. Further, suppose that for $i = 1, \ldots, n$, $Y_i := g_i(X_i)$ for some function $g_i$. Then $\{Y_i\}_{i=1}^n$ is mutually independent.

Proof. It suffices to prove the theorem for $n = 2$. The general case follows easily by induction, using Theorem 8. 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

$$P[(Y_1 = t_1) \cap (Y_2 = t_2)] = P\left[\left(\bigcup_{s_1 \in S'_1} (X_1 = s_1)\right) \cap \left(\bigcup_{s_2 \in S'_2} (X_2 = s_2)\right)\right]$$

$$= \sum_{s_1 \in S'_1} \sum_{s_2 \in S'_2} P[X_1 = s_1] P[X_2 = s_2]$$

$$= \left(\sum_{s_1 \in S'_1} P[X_1 = s_1]\right) \left(\sum_{s_2 \in S'_2} P[X_2 = s_2]\right)$$

$$= P\left[\bigcup_{s_1 \in S'_1} (X_1 = s_1)\right] P\left[\bigcup_{s_2 \in S'_2} (X_2 = s_2)\right] = P[Y_1 = t_1] P[Y_2 = t_2].$$

□

As a special case of the above theorem, if each $g_i$ is the characteristic function for some subset $S'_i$ of the image of $X_i$, then $X_1 \in S'_1, \ldots, X_n \in S'_n$ form a mutually independent family of events.

Exercise 10 Suppose $X$ and $X'$ are random variables that take values in a set $S$ and that have essentially the same distribution. Show that if $f : S \to T$ is a function, then $f(X)$ and $f(X')$ have essentially the same distribution.

Exercise 11 Let $\{X_i\}_{i=1}^n$ be a family of random variables, and let $S_i$ be the image of $X_i$ for $i = 1, \ldots, n$. Show that $\{X_i\}_{i=1}^n$ is mutually independent if and only if for each $i = 2, \ldots, n$, and for all $s_1 \in S_1, \ldots, s_i \in S_i$, we have

$$P[X_i = s_i | (X_1 = s_1) \cap \cdots \cap (X_{i-1} = s_{i-1})] = P[X_i = s_i].$$

Exercise 12 Suppose $X$ and $Y$ are random variables, where $X$ takes values in $S$, and $Y$ takes values in $T$. Further suppose that $Y'$ is uniformly distributed over $T$, and that $(X, Y)$ and $Y'$ are independent. Let $\phi$ be a predicate on $S \times T$. Show that $P[\phi(X, Y) \cap (Y = Y')] = P[\phi(X, Y)]/|T|.$

Exercise 13 Let $X$ and $Y$ be independent random variables, where $X$ is uniformly distributed over a set $S$, and $Y$ is uniformly distributed over a set $T \subset S$. Define a third random variable $Z$ as follows: if $X \in T$, then $Z := X$; otherwise, $Z := Y$. Show that $Z$ is uniformly distributed over $T$. 

17
4 Expectation and variance

Let $P$ be a probability distribution on a sample space $\Omega$. If $X$ is a real-valued random variable, then its expected value, or expectation, is

$$E[X] := \sum_{\omega \in \Omega} X(\omega) P(\omega). \quad (16)$$

If $S$ is the image of $X$, and if for each $s \in S$ we group together the terms in (16) with $X(\omega) = s$, then we see that

$$E[X] = \sum_{s \in S} s P[X = s]. \quad (17)$$

From (17), it is clear that $E[X]$ depends only on the distribution of $X$: if $X'$ is another random variable with the same (or essentially the same) distribution as $X$, then $E[X] = E[X']$.

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 (16) with $X(\omega) = s$, we see that

$$E[f(X)] = \sum_{s \in S} f(s) P[X = s]. \quad (18)$$

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., $X(\omega) = c$ for every $\omega \in \Omega$), then $E[X] = E[c] = c$. Second, if $X$ and $Y$ are random variables such that $X \geq Y$ (i.e., $X(\omega) \geq Y(\omega)$ for every $\omega \in \Omega$), then $E[X] \geq E[Y]$. Similarly, if $X > Y$, then $E[X] > E[Y]$.

In calculating expectations, one rarely makes direct use of (16), (17), or (18), except in rather trivial situations. The next two theorems develop tools that are often quite effective in calculating expectations.

**Theorem 12 (Linearity of expectation)** If $X$ and $Y$ are real-valued random variables, and $a$ is a real number, then

$$E[X + Y] = E[X] + E[Y] \quad \text{and} \quad E[aX] = a E[X].$$

**Proof.** It is easiest to prove this using the defining equation (16) for expectation. For $\omega \in \Omega$, the value of the random variable $X + Y$ at $\omega$ is by definition $X(\omega) + Y(\omega)$, and so we have

$$E[X + Y] = \sum_{\omega} (X(\omega) + Y(\omega)) P(\omega)$$

$$= \sum_{\omega} X(\omega) P(\omega) + \sum_{\omega} Y(\omega) P(\omega)$$

$$= E[X] + E[Y].$$

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

$$E[aX] = \sum_{\omega} (aX(\omega)) P(\omega) = a \sum_{\omega} X(\omega) P(\omega) = a E[X].$$

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

\[
E \left[ \sum_{i \in I} X_i \right] = \sum_{i \in I} E[X_i].
\] (19)

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

**Theorem 13** If \( X \) and \( Y \) are independent, real-valued random variables, then

\[
E[XY] = E[X]E[Y].
\]

**Proof.** It is easiest to prove this using (18), with the function \( f(s, t) := st \) applied to the random variable \((X, Y)\). We have

\[
E[XY] = \sum_{s, t} st P(X = s) \cap (Y = t) = \sum_{s, t} st P(X = s)P(Y = t) = \left( \sum_s s P(X = s) \right) \left( \sum_t t P(Y = t) \right) = E[X]E[Y].
\]

\( \square \)

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

\[
E \left[ \prod_{i \in I} X_i \right] = \prod_{i \in I} E[X_i].
\] (20)

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

**Theorem 14** Let \( X \) be a 0/1-valued random variable. Then

\[
E[X] = P(X = 1).
\]

**Proof.** \( E[X] = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = P(X = 1). \) \( \square \)

**Theorem 15** If \( X \) is a random variable that takes only non-negative integer values, then

\[
E[X] = \sum_{i \geq 1} P(X \geq i).
\]

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

**Proof.** Suppose that the image of \( X \) is contained in \( \{0, \ldots, n\} \), and for \( i = 1, \ldots, n \), let \( X_i \) be the indicator variable for the event \( X \geq i \). Then \( X = X_1 + \cdots + X_n \), and by linearity of expectation and Theorem 14, we have

\[
E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n P(X \geq i).
\] \( \square \)
Let $X$ be a real-valued random variable with $\mu := E[X]$. The **variance** of $X$ is $\text{Var}[X] := E[(X - \mu)^2]$. The variance provides a measure of the spread or dispersion of the distribution of $X$ around its expected value. Note that since $(X - \mu)^2$ takes only non-negative values, variance is always non-negative.

**Theorem 16** Let $X$ be a real-valued random variable, with $\mu := E[X]$, and let $a$ and $b$ be real numbers. Then we have

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

(ii) $\text{Var}[aX] = a^2 \text{Var}[X],$ and

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

**Proof.** For part (i), observe that

$$\text{Var}[X] = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + E[\mu^2] = E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2,$$

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

For part (ii), observe that

$$\text{Var}[aX] = E[a^2X^2] - E[aX]^2 = a^2 E[X^2] - (a\mu)^2$$

$$= a^2 (E[X^2] - \mu^2) = a^2 \text{Var}[X],$$

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:

$$\text{Var}[X + b] = E[(X + b)^2] - (\mu + b)^2$$

$$= (E[X^2] + 2b\mu + b^2) - (\mu^2 + 2b\mu + b^2)$$

$$= E[X^2] - \mu^2 = \text{Var}[X].$$

\[\square\]

The following is an immediate consequence of part (i) of Theorem 16, and the fact that variance is always non-negative:

**Theorem 17** If $X$ is a real-valued random variable, then $E[X^2] \geq E[X]^2$.

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 18** If $\{X_i\}_{i \in I}$ is a finite, pairwise independent family of real-valued random variables, then

$$\text{Var} \left[ \sum_{i \in I} X_i \right] = \sum_{i \in I} \text{Var}[X_i].$$
Proof. We have

\[
\text{Var} \left( \sum_{i \in I} X_i \right) = E\left[ \left( \sum_{i \in I} X_i \right)^2 \right] - \left( E\left[ \sum_{i \in I} X_i \right] \right)^2 \\
= \sum_{i \in I} E[X_i^2] + \sum_{i,j \in I, i \neq j} \left( E[X_i X_j] - E[X_i] E[X_j] \right) - \sum_{i \in I} E[X_i]^2 \\
\text{(by linearity of expectation and rearranging terms)} \\
= \sum_{i \in I} E[X_i^2] - \sum_{i \in I} E[X_i]^2 \\
\text{(by pairwise independence and Theorem 13)} \\
= \sum_{i \in I} \text{Var}[X_i].
\]

□

Corresponding to Theorem 14, we have:

**Theorem 19** Let \( X \) be a 0/1-valued random variable, with \( p := P[X = 1] \) and \( q := P[X = 0] = 1 - p \). Then \( \text{Var}[X] = pq \).

**Proof.** We have \( E[X] = p \) and \( E[X^2] = P[X^2 = 1] = P[X = 1] = p \). Therefore,

\[
\]

□

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

\[
E[X \mid B] = \sum_{\omega \in \Omega} X(\omega) P(\omega \mid B) = P[B]^{-1} \sum_{\omega \in B} X(\omega) P(\omega).
\]

Analogous to (17), if \( S \) is the image of \( X \), we have

\[
E[X \mid B] = \sum_{s \in S} s P[X = s \mid B]. \tag{21}
\]

Furthermore, suppose \( I \) is a finite index set, and \( \{B_i\}_{i \in I} \) is a partition of the sample space, where each \( B_i \) occurs with non-zero probability. If for each \( i \in I \) we group together the terms in (16) with \( \omega \in B_i \), we obtain the **law of total expectation**:

\[
E[X] = \sum_{i \in I} E[X \mid B_i] P[B_i]. \tag{22}
\]

**Example 23** Let \( X \) be uniformly distributed over \( \{1, \ldots, m\} \). Let us compute \( E[X] \) and \( \text{Var}[X] \). We have

\[
E[X] = \sum_{s=1}^{m} s \cdot \frac{1}{m} = \frac{m(m + 1)}{2} \cdot \frac{1}{m} = \frac{m + 1}{2}.
\]
We also have
\[ E[X^2] = \sum_{s=1}^{m} s^2 \cdot \frac{1}{m} = \frac{m(m+1)(2m+1)}{6} \cdot \frac{1}{m} = \frac{(m+1)(2m+1)}{6}. \]

Therefore,
\[ \text{Var}[X] = E[X^2] - E[X]^2 = \frac{m^2 - 1}{12}. \]

**Example 24** 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
\[ E[X | \mathcal{A}] = \frac{2 + 4 + 6}{3} = 4. \]

Similarly, the conditional distribution of \( X \) given \( \overline{\mathcal{A}} \) is essentially the uniform distribution on \( \{1, 3, 5\} \), and so
\[ E[X | \overline{\mathcal{A}}] = \frac{1 + 3 + 5}{3} = 3. \]

Using the law of total expectation, we can compute the expected value of \( X \) as follows:
\[ E[X] = E[X | \mathcal{A}] P[\mathcal{A}] + E[X | \overline{\mathcal{A}}] P[\overline{\mathcal{A}}] = 4 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = \frac{7}{2}, \]
which agrees with the calculation in the previous example.

**Example 25** Let \( X \) be a random variable with a binomial distribution, as in Example 18, that counts the number of successes among \( n \) Bernoulli trials, each of which succeeds with probability \( p \). Let us compute \( E[X] \) and \( \text{Var}[X] \). We can write \( X \) as the sum of indicator variables, \( X = \sum_{i=1}^{n} X_i \), where \( X_i \) is the indicator variable for the event that the \( i \)th trial succeeds; each \( X_i \) takes the value 1 with probability \( p \) and 0 with probability \( q := 1 - p \), and the family of random variables \( \{X_i\}_{i=1}^{n} \) is mutually independent (see Example 22). By Theorems 14 and 19, we have \( E[X_i] = p \) and \( \text{Var}[X_i] = pq \) for \( i = 1, \ldots, n \). By linearity of expectation, we have
\[ E[X] = \sum_{i=1}^{n} E[X_i] = np. \]

By Theorem 18, and the fact that \( \{X_i\}_{i=1}^{n} \) is mutually independent (and hence pairwise independent), we have
\[ \text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = npq. \]

**Example 26** Our proof of Theorem 1 could be elegantly recast in terms of indicator variables. For \( B \subset \Omega \), let \( X_B \) be the indicator variable for \( B \), so that \( X_B(\omega) = \delta_{\omega}[B] \) for each \( \omega \in \Omega \). Equation (8) then becomes
\[ X_B = \sum_{\emptyset \subsetneq J \subset I} (-1)^{|J|-1} X_{A_J}, \]
and by Theorem 14 and linearity of expectation, we have
\[ P[\mathcal{A}] = E[X_B | \mathcal{A}] = \sum_{\emptyset \subsetneq J \subset I} (-1)^{|J|-1} E[X_{A_J}] = \sum_{\emptyset \subsetneq J \subset I} (-1)^{|J|-1} P[X_{A_J}]. \]
Exercise 14 Suppose $X$ is a real-valued random variable. Show that $|E[X]| \leq E[|X|] \leq E[X^2]^{1/2}$.

Exercise 15 Suppose $X$ and $Y$ take non-negative real values, and that $Y \leq c$ for some constant $c$. Show that $E[XY] \leq cE[X]$.

Exercise 16 Let $X$ be a 0/1-valued random variable. Show that $\text{Var}[X] \leq 1/4$.

Exercise 17 Let $\mathcal{B}$ be an event with $P[\mathcal{B}] \neq 0$, and let $\{\mathcal{B}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union is $\mathcal{B}$. Generalizing the law of total expectation (22), show that for every real-valued random variable $X$, if $I^* := \{i \in I : P[\mathcal{B}_i] \neq 0\}$, then we have

$$E[X \mid \mathcal{B}]P[\mathcal{B}] = \sum_{i \in I^*} E[X \mid \mathcal{B}_i]P[\mathcal{B}_i].$$

Also show that if $E[X \mid \mathcal{B}_i] \leq \alpha$ for each $i \in I^*$, then $E[X \mid \mathcal{B}] \leq \alpha$.

Exercise 18 Let $\mathcal{B}$ be an event with $P[\mathcal{B}] \neq 0$, and let $\{\mathcal{C}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union contains $\mathcal{B}$. Again, generalizing the law of total expectation, show that for every real-valued random variable $X$, if $I^* := \{i \in I : P[\mathcal{B} \cap \mathcal{C}_i] \neq 0\}$, then we have

$$E[X \mid \mathcal{B}] = \sum_{i \in I^*} E[X \mid \mathcal{B} \cap \mathcal{C}_i]P[\mathcal{C}_i \mid \mathcal{B}].$$

Exercise 19 This exercise makes use of the notion of a convex function (see §A6).

(a) Prove Jensen’s inequality: if $f$ is convex on an interval, and $X$ is a random variable taking values in that interval, then $E[f(X)] \geq f(E[X])$. Hint: use induction on the size of the image of $X$. (Note that Theorem 17 is a special case of this, with $f(s) := s^2$.)

(b) Using part (a), show that if $X$ takes non-negative real values, and $\alpha$ is a positive number, then $E[X^\alpha] \geq E[X]^{\alpha}$ if $\alpha \geq 1$, and $E[X^\alpha] \leq E[X]^{\alpha}$ if $\alpha \leq 1$.

(c) Using part (a), show that if $X$ takes positive real values, then $E[X] \geq e^{E[\log X]}$.

(d) Using part (c), derive the arithmetic/geometric mean inequality: for all positive numbers $x_1, \ldots, x_n$, we have

$$(x_1 + \cdots + x_n)/n \geq (x_1 \cdots x_n)^{1/n}.$$

Exercise 20 For real-valued random variables $X$ and $Y$, their covariance is defined as $\text{Cov}[X, Y] := E[XY] - E[X]E[Y]$. Show that:

(a) if $X$, $Y$, and $Z$ are real-valued random variables, and $a$ is a real number, then $\text{Cov}[X + Y, Z] = \text{Cov}[X, Z] + \text{Cov}[Y, Z]$ and $\text{Cov}[aX, Z] = a \text{Cov}[X, Z]$;

(b) if $\{X_i\}_{i \in I}$ is a finite family of real-valued random variables, then

$$\text{Var}\left[\sum_{i \in I} X_i\right] = \sum_{i \in I} \text{Var}[X_i] + \sum_{i,j \in I, i \neq j} \text{Cov}[X_i, X_j].$$

Exercise 21 Consider again the game played between Alice and Bob in Example 11. Suppose that to play the game, Bob must place a one dollar bet. However, after Alice reveals the sum of the two dice, Bob may elect to double his bet. If Bob’s guess is correct, Alice pays him his bet, and otherwise Bob pays Alice his bet. Describe an optimal playing strategy for Bob, and calculate his expected winnings.

Exercise 22 A die is rolled repeatedly until it comes up “1,” or until it is rolled $n$ times (whichever comes first). What is the expected number of rolls of the die?
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 20 (Markov’s inequality)** Let $X$ be a random variable that takes only non-negative real values. Then for every $\alpha > 0$, we have

$$P[X \geq \alpha] \leq \frac{E[X]}{\alpha}.$$  

**Proof.** We have

$$E[X] = \sum_s s \cdot P[X = s] = \sum_{s < \alpha} s \cdot P[X = s] + \sum_{s \geq \alpha} s \cdot P[X = s],$$

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,

$$E[X] \geq \sum_{s \geq \alpha} s \cdot P[X = s] \geq \sum_{s \geq \alpha} \alpha \cdot P[X = s] = \alpha \cdot P[X \geq \alpha].$$

□

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

$$P[|X - \mu| \geq \alpha] \leq \frac{\nu}{\alpha^2}.$$  

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

$$P[|X - \mu| \geq \alpha] = P[Y \geq \alpha^2] \leq \frac{\nu}{\alpha^2}.$$  

□

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

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

The variable $\overline{X}$ is called the **sample mean** of $\{X_i\}_{i \in I}$. By the linearity of expectation, we have $E[\overline{X}] = \mu$, and since $\{X_i\}_{i \in I}$ is pairwise independent, it follows from Theorem 18 (along with part (ii) of Theorem 16) that $\text{Var}[\overline{X}] = \nu/n$. Applying Chebyshev’s inequality, for every $\epsilon > 0$, we have

$$P[|\overline{X} - \mu| \geq \epsilon] \leq \frac{\nu}{n \epsilon^2}. \quad (23)$$

The inequality (23) says that for all $\epsilon > 0$, and for all $\delta > 0$, there exists $n_0$ (depending on $\epsilon$ and $\delta$, as well as the variance $\nu$) such that $n \geq n_0$ implies

$$P[|\overline{X} - \mu| \geq \epsilon] \leq \delta. \quad (24)$$

In words:
As $n$ gets large, the sample mean closely approximates the expected value $\mu$ with high probability.

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

Let us now examine an even more specialized case of the above situation, where each $X_i$ is a 0/1-valued random variable, taking the value 1 with probability $p$, and 0 with probability $q := 1 - p$. By Theorems 14 and 19, the $X_i$'s have a common expected value $p$ and variance $pq$. Therefore, by (23), for every $\epsilon > 0$, we have

$$P[|\bar{X} - p| \geq \epsilon] \leq \frac{pq}{n\epsilon^2}. \quad (25)$$

The bound on the right-hand side of (25) decreases linearly in $n$. If one makes the stronger assumption that the family $\{X_i\}_{i \in I}$ is mutually independent (so that $X := \sum_{i} X_i$ has a binomial distribution), one can obtain a much better bound that decreases exponentially in $n$:

**Theorem 22 (Chernoff bound)** Let $\{X_i\}_{i \in I}$ be a finite, non-empty, and mutually independent family of random variables, such that each $X_i$ is 1 with probability $p$ and 0 with probability $q := 1 - p$. Assume that $0 < p < 1$. Also, let $n := |I|$ and $\bar{X}$ be the sample mean of $\{X_i\}_{i \in I}$. Then for every $\epsilon > 0$, we have:

1. $P[\bar{X} - p \geq \epsilon] \leq e^{-n\epsilon^2/2q}$;
2. $P[\bar{X} - p \leq -\epsilon] \leq e^{-n\epsilon^2/2p}$;
3. $P[|\bar{X} - p| \geq \epsilon] \leq 2e^{-n\epsilon^2/2}$.

**Proof.** First, we observe that (ii) follows directly from (i) by replacing $X_i$ by $1 - X_i$ and exchanging the roles of $p$ and $q$. Second, we observe that (iii) follows directly from (i) and (ii). Thus, it suffices to prove (i).

Let $\alpha > 0$ be a parameter, whose value will be determined later. Define the random variable $Z := e^{\alpha(\bar{X} - p)}$. Since the function $x \mapsto e^{\alpha x}$ is strictly increasing, we have $\bar{X} - p \geq \epsilon$ if and only if $Z \geq e^{\alpha\epsilon}$. By Markov’s inequality, it follows that

$$P[\bar{X} - p \geq \epsilon] = P[Z \geq e^{\alpha\epsilon}] \leq E[Z]e^{-\alpha\epsilon}. \quad (26)$$

So our goal is to bound $E[Z]$ from above.

For each $i \in I$, define the random variable $Z_i := e^{\alpha(X_i - p)}$. Observe that $Z = \prod_{i \in I} Z_i$, that $\{Z_i\}_{i \in I}$ is a mutually independent family of random variables (see Theorem 11), and that for each $i \in I$, we have

$$E[Z_i] = e^{\alpha(1-p)}p + e^{\alpha(0-p)}q = pe^{\alpha q} + qe^{-\alpha p}.$$ 

It follows that

$$E[Z] = E[\prod_{i \in I} Z_i] = \prod_{i \in I} E[Z_i] = (pe^{\alpha q} + qe^{-\alpha p})^n.$$ 

We will prove below that

$$pe^{\alpha q} + qe^{-\alpha p} \leq e^{\alpha^2 q/2}. \quad (27)$$

From this, it follows that

$$E[Z] \leq e^{\alpha^2 qn/2}. \quad (28)$$

Combining (28) with (26), we obtain

$$P[\bar{X} - p \geq \epsilon] \leq e^{\alpha^2 qn/2 - \alpha\epsilon}. \quad (29)$$
Now we choose the parameter $\alpha$ so as to minimize the quantity $\alpha^2 n/2 - \alpha n e$. The optimal value of $\alpha$ is easily seen to be $\alpha = \epsilon/q$, and substituting this value of $\alpha$ into (29) yields (i).

To finish the proof of the theorem, it remains to prove the inequality (27). Let

$$\beta := pe^{\alpha q} + q e^{-\alpha p}.$$ 

We want to show that $\beta \leq e^{\alpha^2 q/2}$, or equivalently, that $\log \beta \leq \alpha^2 q/2$. We have

$$\beta = e^{\alpha q}(p + q e^{-\alpha}) = e^{\alpha q}(1 - q(1 - e^{-\alpha})).$$

Taking logarithms, and applying the inequalities in §A1, we obtain

$$\log \beta = \alpha q + \log(1 - q(1 - e^{-\alpha})) \leq \alpha q - q(1 - e^{-\alpha}) = q(e^{-\alpha} + \alpha - 1) \leq q\alpha^2/2.$$ 

This establishes (27) and completes the proof of the theorem. \(\square\)

Thus, the Chernoff bound is a quantitatively superior version of the law of large numbers, although its range of application is clearly more limited.

**Example 27** Suppose we toss a fair coin 10,000 times. The expected number of heads is 5,000. What is an upper bound on the probability $\alpha$ that we get 6,000 or more heads? Using Markov’s inequality, we get $\alpha \leq 5/6$. Using Chebyshev’s inequality, and in particular, the inequality (25), we get

$$\alpha \leq \frac{1/4}{10^4 10^{-2}} = \frac{1}{400}.$$ 

Finally, using the Chernoff bound, we obtain

$$\alpha \leq e^{-10^4 10^{-2}/2(0.5)} = e^{-100} \approx 10^{-43.4}.$$ 

**Exercise 23** With notation and assumptions as in Theorem 22, and with $p := q := 1/2$, show that there exist constants $c_1$ and $c_2$ such that

$$P[|X - 1/2| \geq c_1 / \sqrt{n}] \leq 1/4 \quad \text{and} \quad P[|X - 1/2| \geq c_2 / \sqrt{n}] \geq 1/4.$$ 

**Hint:** Stirling’s approximation may be useful here.

**Exercise 24** In each step of a random walk, we toss a coin, and move either one unit to the right, or one unit to the left, depending on the outcome of the coin toss. The question is, after $n$ steps, what is our expected distance from the starting point? Let us model this using a mutually independent family of random variables $\{Y_i\}_{i=1}^n$, with each $Y_i$ uniformly distributed over $\{-1, 1\}$, and define $Y := Y_1 + \cdots + Y_n$. Show that the $c_1 \sqrt{n} \leq |E[Y]| \leq c_2 \sqrt{n}$, for some constants $c_1$ and $c_2$.

**Exercise 25** You are given three biased coins, where for $i = 1, 2, 3$, coin $i$ comes up heads with probability $p_i$. The coins look identical, and all you know is the following: (1) $|p_1 - p_2| > 0.01$ and (2) either $p_3 = p_1$ or $p_3 = p_2$. Your goal is to determine whether $p_3$ is equal to $p_1$, or to $p_2$. Design a random experiment to determine this. The experiment may produce an incorrect result, but this should happen with probability at most $10^{-12}$. Try to use a reasonable number of coin tosses.

**Exercise 26** Consider the following game, parameterized by a positive integer $n$. One rolls a pair of dice, and records the value of their sum. This is repeated until some value $\ell$ is recorded $n$ times, and this value $\ell$ is declared the “winner.” It is intuitively clear that 7 is the most likely winner. Let $\alpha_n$ be the probability that 7 does not win. Give a careful argument that $\alpha_n \to 0$ as $n \to \infty$. Assume that the rolls of the dice are mutually independent.
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 \( \Omega \) is countably infinite simply means that there is a bijection \( f \) from the set of positive integers onto \( \Omega \); thus, we can enumerate the elements of \( \Omega \) as \( \omega_1, \omega_2, \omega_3, \ldots \), where \( \omega_i := f(i) \).

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

Example 28 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 \( P \), where the sample space consists of the set of all positive integers: for each positive integer \( k \), \( P(k) := 2^{-k} \). 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 29 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 \( P(k) := q^{k-1}p \) 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 30 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 \( A \) of \( \Omega \). The probability \( P[A] \) of \( A \) is defined as the sum of the probabilities associated with the elements of \( A \). This sum is treated as an infinite series when \( A \) is infinite. This series is guaranteed to converge, and its value does not depend on the particular enumeration of the elements of \( A \).

Example 31 Consider the geometric distribution discussed in Example 29, where \( p \) is the success probability of each Bernoulli trial, and \( q := 1-p \). For a given integer \( i \geq 1 \), consider the event \( A \) that the number of trials executed is at least \( i \). Formally, \( A \) is the set of all integers greater than
or equal to $i$. Intuitively, $P[A]$ 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

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

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

(i) $P[A] \leq \sum_{i=1}^{\infty} P[A_i]$, and

(ii) $P[A] = \sum_{i=1}^{\infty} P[A_i]$ if $\{A_i\}_{i=1}^{\infty}$ is pairwise disjoint.

*Proof.* As in the proof of Theorem 1, for $\omega \in \Omega$ and $B \subset \Omega$, define $\delta_{\omega}(B) := 1$ if $\omega \in B$, and $\delta_{\omega}(B) := 0$ if $\omega \notin B$. First, suppose that $\{A_i\}_{i=1}^{\infty}$ is pairwise disjoint. Evidently, $\delta_{\omega}[A] = \sum_{i=1}^{\infty} \delta_{\omega}[A_i]$ for each $\omega \in \Omega$, and so

$$P[A] = \sum_{\omega \in \Omega} P(\omega) \delta_{\omega}[A] = \sum_{\omega \in \Omega} P(\omega) \sum_{i=1}^{\infty} \delta_{\omega}[A_i]$$

$$= \sum_{i=1}^{\infty} \sum_{\omega \in \Omega} P(\omega) \delta_{\omega}[A_i] = \sum_{i=1}^{\infty} P[A_i],$$

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 $\{A'_i\}_{i=1}^{\infty}$, where $A'_i := A_i \setminus \bigcup_{j=1}^{i-1} A_i$, as $P[A] = \sum_{i=1}^{\infty} P[A'_i] \leq \sum_{i=1}^{\infty} P[A_i]$. □

### 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 (9) and (10)), as well as Bayes’ theorem (11), extend to families of events $\{B_i\}_{i \in I}$ indexed by any countably infinite set $I$. The definitions of independent families of events ($k$-wise and mutually) extend verbatim to infinite families.

### 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 (16); that is, $E[X] := \sum_{\omega} X(\omega) P(\omega)$, 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 $E[X]$ is finite. In this case, the series defining $E[X]$ converges to the same finite limit, regardless of the ordering of the terms.

If $E[X]$ is not finite, then under the right conditions, $E[X]$ 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 $E[X]$ is not
finite, then we naturally define \( E[X] := \infty \), as the series defining \( E[X] \) diverges to \( \infty \), regardless of the ordering of the terms. In the general case, we may define random variables \( X^+ \) and \( X^- \), where

\[
X^+(\omega) := \max\{0, X(\omega)\} \text{ and } X^-(\omega) := \max\{0, -X(\omega)\},
\]

so that \( X = X^+ - X^- \), 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 \( E[X] \) is not finite, so that one of \( E[X^+] \) or \( E[X^-] \) is infinite. If \( E[X^+] = E[X^-] = \infty \), then we say that \( E[X] \) does not exist; otherwise, we define \( E[X] := E[X^+] - E[X^-] \), which is \( \pm \infty \); in this case, the series defining \( E[X] \) diverges to \( \pm \infty \), regardless of the ordering of the terms.

**Example 32** Let \( X \) be a random variable whose distribution is as in Example 30. Since the series \( \sum_{k=1}^{\infty} 1/k^2 \) converges and the series \( \sum_{k=1}^{\infty} 1/k \) diverges, the expectation \( E[X] \) is finite, while \( E[X^2] = \infty \). One may also verify that the random variable \((-1)^X X^2 \) has no expectation.

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

If \( E[X] \) exists, then we can regroup the terms in the series \( \sum_\omega X(\omega) P(\omega) \), without affecting its value. In particular, equation (17) holds provided \( E[X] \) exists, and equation (18) holds provided \( E[f(X)] \) exists.

Theorem 12 still holds, under the additional hypothesis that \( E[X] \) and \( E[Y] \) are finite. Equation (19) also holds, provided the individual expectations \( E[X_i] \) are finite. More generally, if \( E[X] \) and \( E[Y] \) exist, then \( E[X + Y] = E[X] + E[Y] \), unless \( E[X] = \infty \) and \( E[Y] = -\infty \), or \( E[X] = -\infty \) and \( E[Y] = \infty \). Also, if \( E[X] \) exists, then \( E[aX] = a E[X] \), unless \( a = 0 \) and \( E[X] = \pm \infty \).

One might consider generalizing (19) to countably infinite families of random variables. To this end, suppose \( \{X_i\}_{i=1}^{\infty} \) is an infinite sequence of real-valued random variables. The random variable \( X := \sum_{i=1}^{\infty} X_i \) is well defined, provided the series \( \sum_{i=1}^{\infty} X_i(\omega) \) converges for each \( \omega \in \Omega \). One might hope that \( E[X] = \sum_{i=1}^{\infty} E[X_i] \); however, this is not in general true, even if the individual expectations, \( E[X_i] \), are non-negative, and even if the series defining \( X \) converges absolutely for each \( \omega \); nevertheless, it is true when the \( X_i \)'s are non-negative:

**Theorem 24** Let \( \{X_i\}_{i=1}^{\infty} \) 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 \( \sum_{i=1}^{\infty} X_i(\omega) \) converges for each \( \omega \in \Omega \), and define \( X := \sum_{i=1}^{\infty} X_i \). Then we have

\[
E[X] = \sum_{i=1}^{\infty} E[X_i].
\]

**Proof.** This is a calculation just like the one made in the proof of Theorem 23, where, again, we use the fact that we may reverse the order of summation in an infinite double summation of non-negative terms:

\[
E[X] = \sum_{\omega \in \Omega} P(\omega) X(\omega) = \sum_{\omega \in \Omega} P(\omega) \sum_{i=1}^{\infty} X_i(\omega)
= \sum_{i=1}^{\infty} \sum_{\omega \in \Omega} P(\omega) X_i(\omega) = \sum_{i=1}^{\infty} E[X_i].
\]

\[\square\]
Theorem 13 holds under the additional hypothesis that $E[X]$ and $E[Y]$ are finite. Equation (20) also holds, provided the individual expectations $E[X_i]$ are finite. Theorem 15 also holds, but where the sum may be infinite; it can be proved using essentially the same argument as in the finite case, combined with Theorem 24.

**Example 33** Suppose $X$ is a random variable with a geometric distribution, as in Example 29, with an associated success probability $p$ and failure probability $q := 1 − p$. As we saw in Example 31, for every integer $i ≥ 1$, we have $P[X ≥ i] = q^{i−1}$. We may therefore apply the infinite version of Theorem 15 to easily compute the expected value of $X$:

$$E[X] = \sum_{i=1}^{∞} P[X ≥ i] = \sum_{i=1}^{∞} q^{i−1} = \frac{1}{1−q} = \frac{1}{p}.$$  

**Example 34** To illustrate that Theorem 24 does not hold in general, consider the geometric distribution on the positive integers, where $P(j) = 2^{−j}$ for $j ≥ 1$. For $i ≥ 1$, define the random variable $X_i$ so that $X_i(i) = 2^i$, $X_i(i+1) = −2^{i+1}$, and $X_i(j) = 0$ for all $j \notin \{i, i+1\}$. Then $E[X_i] = 0$ for all $i ≥ 1$, and so $\sum_{i=1}^{∞} E[X_i] = 0$. Now define $X := \sum_{i=1}^{∞} X_i$. This is well defined, and in fact $X(1) = 2$, while $X(j) = 0$ for all $j > 1$. Hence $E[X] = 1$.

The variance $\text{Var}[X]$ of $X$ exists only when $µ := E[X]$ is finite, in which case it is defined as usual as $E[(X − µ)^2]$, which may be either finite or infinite. Theorems 16, 17, and 18 hold provided all the relevant expectations and variances are finite.

The definition of conditional expectation carries over verbatim. Equation (21) holds, provided $E[X | B]$ exists, and the law of total expectation (22) holds, provided $E[X]$ exists. The law of total expectation also holds for a countably infinite partition $\{B_i\}_{i \in I}$, provided $E[X]$ exists, and each of the conditional expectations $E[X | B_i]$ is finite.

### 6.5 Some useful bounds

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

**Exercise 27** Let $\{A_i\}_{i=1}^{∞}$ be a family of events, such that $A_i \subset A_{i+1}$ for each $i ≥ 1$, and let $A := \bigcup_{i=1}^{∞} A_i$. Show that $P[A] = \lim_{i \to ∞} P[A_i]$.

**Exercise 28** Generalize Exercises 5, 6, 17, and 18 to the discrete setting, allowing a countably infinite index set $I$.

**Exercise 29** Suppose $X$ is a random variable taking positive integer values, and that for some real number $q$, with $0 ≤ q ≤ 1$, and for all integers $i ≥ 1$, we have $P[X ≥ i] = q^{i−1}$. Show that $X$ has a geometric distribution with associated success probability $p := 1 − q$.

**Exercise 30** This exercise extends Jensen's inequality (see Exercise 19) to the discrete setting. Suppose that $f$ is a convex function on an interval $I$. Let $X$ be a random variable whose image is a countably infinite subset of $I$, and assume that both $E[X]$ and $E[f(X)]$ are finite. Show that $E[f(X)] ≥ f(E[X])$. Hint: use continuity.

**Exercise 31** A gambler plays a simple game in a casino: with each play of the game, the gambler may bet any number $m$ of dollars; a fair coin is tossed, and if it comes up heads, the casino pays $m$ dollars to the gambler, and otherwise, the gambler pays $m$ dollars to the casino. The gambler plays
the game repeatedly, using the following strategy: he initially bets a dollar, and with each subsequent play, he doubles his bet; if he ever wins, he quits and goes home; if he runs out of money, he also goes home; otherwise, he plays again. Show that if the gambler has an infinite amount of money, then his expected winnings are one dollar, and if he has a finite amount of money, his expected winnings are zero.

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
\[
\binom{n}{k} < \binom{n}{k+1} \quad \text{for } 0 \leq k < (n-1)/2, \\
\binom{n}{k} > \binom{n}{k+1} \quad \text{for } (n-1)/2 < k < n, \text{ and} \\
\binom{n}{k} = \binom{n}{n-k} \quad \text{for } 0 \leq k \leq n.
\]

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 \( \bigcup_{i=0}^{\infty} \) 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 \( \sum_{k} x_{ik_j} \). We then have \( \sum_{i} \sum_{j} x_{ij} = \sum_{j} \sum_{i} x_{ij} = \sum_{k} x_{ik_j} \), 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 \( \sum_k |x_{ik}j_k| < \infty \).

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 \( \sum_k a_{ik}b_{jk} \) 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 \( I \).