as specified by the conditions of the theorem. Then $s(p)$ satisfies condition C_{t-1}. Indeed, if $s(p)$ contained a pattern q from P_{t-1}, then it follows from Proposition 14.2 that p would have to contain a pattern from P_t. (There had to be something large between the entries playing the role of $t+1$ and 1 in q.) Therefore, $s(p)$ is $(t-1)$-stack sortable by the induction hypothesis, so p is t-stack sortable.

(20) We claim that there are no such permutations. We know by Lemma 14.5 that $s(p) = s(f(p))$, where f is the map given by Definition 14.8. On the other hand, Proposition 14.4 shows that $d(p)+d(f(p)) = n-1$. Therefore, if n is even, then one of p and $f(p)$ must have an odd number of descents, and the other one must have an even number of descents. So $p \neq f(p)$, while $s(p) = s(f(p))$.

(21) No, that is not true. A counterexample is 163452. This permutation is not 2-stack sortable because of the 2341-pattern 3452. The "only if" part is true. If there are at least two entries on the left of n that are larger than the entry c located on the right of n, then let a and b be the *leftmost* two entries with this property. If $a < b$, then $abnc$ is a 2341-pattern, and if $b < a$, then $abnc$ is a 3241-pattern that is not part of a 35241-pattern. (There is nothing between a and b that is larger than c.)

Chapter 15

Who Knows What It Looks Like, But It Exists. The Probabilistic Method

We use the words "likely" or "probable" and "likelihood" or "probability" every day in informal conversations. While making these concepts absolutely rigorous can be a difficult task, we will concentrate on special cases when a mathematical definition of probability is straightforward, and conforms to common sense.

15.1 The Notion of Probability

Assume we toss a coin four times, and want to know the probability that we will get at least three heads. It is clear that the number of all outcomes of the four coin tosses is $2^4 = 16$. Indeed, each coin toss can result in two possible outcomes. On the other hand, the number of *favorable* outcomes of our coin tossing sequence is five. Indeed, the five favorable outcomes, that is, those containing at least three heads, are $HHHH, HHHT, HHTH, HTHH$, and $THHH$. Our common sense now suggests that we define the probability of getting at least three heads as the ratio of the number of favorable outcomes to the number of all outcomes. Doing that, we get that the probability of getting at least three heads is 5/16.

This common sense approach is the basis of our formal definition of probability. It goes without saying that we will have to be a little more careful. For instance, the above argument assumed, without mentioning it, that our coin is fair, that is, a coin toss is equally likely to result in a head or tail.

Definition 15.1 Let Ω be a finite set of outcomes of some sequence of trials, so that all these outcomes are equally likely. Let $A \subseteq \Omega$. Then Ω is called a *sample space*, and A is called an *event*. The ratio

$$P(A) = \frac{|A|}{|\Omega|}$$

is called the probability of A.

In particular, P is a function that is defined on the set of all subsets of Ω, and $0 \leq P(A) \leq 1$ always holds.

There are, of course, circumstances when this definition does not help, namely when Ω and A are not finite sets. An example of that situation is to compute the probability that a randomly thrown ball hits a given tree. As the ball could be thrown in infinitely many directions, and would hit the tree in an infinite number of cases, the above definition would be useless. We will not discuss that situation in this book; we will only study finite sample spaces.

Note that if A and B are disjoint subsets of Ω, then we have $|A \cup B| = |A| + |B|$, and therefore, $P(A \cup B) = P(A) + P(B)$. In general, we know from the Sieve formula that $|A \cup B| = |A| + |B| - |A \cap B|$, implying $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. A generalization of this observation is the following simple, but extremely useful inequality.

Proposition 15.1 *Let $A_1, A_2, \cdots, A_n$ be events from the same sample space. Then we have*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n).$$

Proof. We simply have to show that

$$|A_1 \cup \cdots \cup A_n| \leq |A_1| + \cdots + |A_n|.$$

This is true as the left-hand side counts each element of the sample space that is part of at least one of the A_i *exactly once*, while the right-hand side counts each element of the sample space that is part of at least one of the A_i *at least once*. □

The reader has already been subjected to some training in basic enumeration in Chapters 3–7. Most exercises in those chapters can be formulated in the language of probability. For example, the question "how many six-digit integers contain the digit 6" can be asked as "what is the probability that a randomly chosen six-digit integer contains the digit 6". Therefore, we do not cover these basic questions again here. Instead, we close this section by two examples that show how *counterintuitive* probabilities can be.

Example 15.1 In one of the lottery games available in Florida, six numbers are drawn from the set of numbers $1, 2, \cdots, 36$. What is the probability that a randomly selected ticket will contain at least one winning number?

Some people tend to answer $\frac{6}{36} = \frac{1}{6}$ to this question. They are wrong. That answer would be correct if *only one* number were drawn. Then the number of favorable outcomes would indeed be six, and the number of all outcomes would indeed be 36. However, when six numbers are drawn, the situation is more complicated.

Proof. Let A be the event that a ticket contains at least one winning number, and let B be the event that a ticket does not contain any winning number. Then clearly, A and B are disjoint, and $A \cup B = \Omega$, so $P(A) + P(B) = 1$. Therefore, it suffices to compute $P(B)$. For a ticket not to contain any winning numbers, it has to contain six non-winning numbers. The number of ways that can happen is $\binom{30}{6}$. Therefore,

$$P(A) = 1 - P(B) = 1 - \frac{\binom{30}{6}}{\binom{36}{6}} = 1 - 0.3048 = 0.6952.$$

So with almost 70 percent probability, a randomly chosen ticket will contain at least one winning number! No wonder you must have more than one winning number to actually win a prize. □

Note that when A and B are two disjoint events, then we say that A and B are *mutually exclusive*. In other words, it is not possible that A and B happen together. If, in addition, we also have $A \cup B = \Omega$, then we say that B is the *complement* of A. We denote this by writing $\bar{A} = B$.

Example 15.2 Forty people are present at a party, and there is nobody among them who was born on February 29. Adam proposes the following game to Bill. Each guest writes his or her birthday (just day and month, not the year) on a piece of paper. If there are two pieces of paper with the same date on them, then Adam wins, if not, then Bill. When Bill heard this proposal, he looked around, and said "Fine, there are only forty people here, much less than the number of days in a year, so I am bound to win." What do we think about Bill's argument?

Proof. The problem with Bill's argument is that he fails to note the difference between one hundred percent probability and more than fifty percent probability. If we want to be one hundred percent sure that there will be two people in the room having the same birthday, then we would indeed need 366 people to be present. To have more than fifty percent chance is an entirely different issue.

In what follows, we prove that if there are at least 23 people at the party, then Adam, not Bill, has more chance of winning this game. In order to prove this, it is clearly sufficient to provide a proof for the case when there are exactly 23 people at the party as any additional person just improves Adam's chances.

Let us compute the probability that there are *no two people* at the party who have the same birthday. For that to happen, the first person's birthday can be any of the 365 possible days of the year, that of the second person could be any of 364 days, and so on. So the number of favorable outcomes is $(365)_{23}$. On the other hand, the number of all outcomes is obviously 365^{23}. Therefore, the probability that there are no two people in the room whose birthdays coincide is

$$\frac{365 \cdot 364 \cdots 343}{365^{23}} = \frac{364 \cdot 363 \cdots 343}{365^{22}} < \frac{1}{2}.$$

Therefore, the probability that there are two people at the party who *do* have the same birthday is *more* than one half. □

Finally, we point out that the condition that nobody was born on February 29 was only included to make the situation simpler. Indeed, February 29 exists only in leap-years, so the chance of being born on that day is 1/4 of the chance of being born on any other given day. That would make the outcomes in our sample space not equally likely, contradicting the definition of sample space. We could help this by changing our sample space from the 365-element set of dates in a year to the set of $4 \cdot 365 + 1 = 1461$ days of a 4-year cycle. That would make computations a little more cumbersome.

15.2 Nonconstructive Proofs

If there are balls in a box, and we know that the probability that a randomly selected ball is blue is more than 0, then we can certainly conclude that there is at least one blue ball in the box. This thought seems utterly simple at first sight, but it has proved to be extremely useful in existence proofs as the following examples show.

Recall that in Chapter 13, we defined the symmetric Ramsey number $R(k,k)$. For easy reference, this was the smallest positive integer so that if we 2-color the edges of the complete graph on $R(k,k)$ vertices, we always get a K_k subgraph whose edges are all the same color.

Let us try to find a lower bound for $R(k,k)$ by proving that $R(k,k) > 2^{k/2}$. Let us take a closer look at this statement. What it says is that if G is a complete graph on $2^{k/2}$ vertices, then *it is possible* to 2-color the edges of G so that no monochromatic copy of K_k is formed. When we proved similar statements in Chapter 13, showing that $R(3,3) > 5$, or $R(4,4) > 17$, we proved them by actually providing a coloring of K_5 or K_{17} that indeed did not contain the required monochromatic copies. However, this was more than what we strictly needed to do. To prove $R(k,k) > 2^{k/2}$, it suffices to prove that it is *possible* to 2-color the edges of G so that no monochromatic copy of K_k is formed; it is *not* necessary to *actually find* such a coloring. We will shortly see how big a difference this is.

Theorem 15.1 *For all positive integers $k \geq 3$, we have $R(k,k) > 2^{k/2}$.*

Proof. Let $G = K_n$, and let us color each edge of G red or blue as follows. For each edge, we flip a coin. If we get a head, we color that edge red, otherwise we color that edge blue. This way each edge will be red with probability one half, and blue with probability one half. We are going to show that the probability p that we get no monochromatic K_k-subgraphs in G this way is more than zero. On the other hand, $p = \frac{|F|}{|\Omega|}$, the number of favorable outcomes divided by the number of all outcomes, where Ω is the set of all possible 2-colorings of the edges of a complete graph on n vertices.

So $p > 0$ implies that there is at least one favorable outcome, that is, there is at least one K_n with 2-colored edges that does not contain any monochromatic K_k-subgraphs.

Instead of proving that $p > 0$, we will prove that $1 - p < 1$, which is clearly an equivalent statement. Note that $1 - p$ is the probability that we get at least one monochromatic subgraph in our randomly colored graph

$G = K_n$.

The number of ways to 2-color the edges of a given K_k-subgraph of K_n is clearly $2^{\binom{k}{2}}$ as we have two choices for the color of each edge. Out of all these colorings, only two will be monochromatic, one with all edges red, and one with all edges blue. Therefore, the probability that a randomly chosen K_k-subgraph is monochromatic is

$$\frac{2}{2^{\binom{k}{2}}} = 2^{1-\binom{k}{2}}.$$

The graph K_n has $\binom{n}{k}$ subgraphs that are isomorphic to K_k. Obviously, each of them has the same chance to be monochromatic. On the other hand, the probability that *at least one* of them is monochromatic is *at most* the sum of these $\binom{n}{k}$ individual probabilities, by Proposition 15.1. In other words, if A_S denotes the event that the K_k-subgraph S of G has monochromatic edges, then

$$P(\cup_S A_S) \leq \sum_S P(A_S) = \binom{n}{k} 2^{1-\binom{k}{2}}, \qquad (15.1)$$

where S ranges through all K_k-subgraphs of G. Now assume, in accordance with our criterion, that $n \leq 2^{k/2}$. Then the last term of (15.1) can be bounded as follows.

$$\binom{n}{k} 2^{1-\binom{k}{2}} < \frac{n^k}{k!} \cdot 2^{1-\binom{k}{2}} \leq \frac{2 \cdot 2^{k^2/2}}{k! 2^{\binom{k}{2}}} = 2\frac{2^{k/2}}{k!} < 1,$$

for all $k \geq 3$. The last inequality is very easy to prove, for example by induction. □

We have seen in Chapter 13 that $R(k,k) \leq 4^k$. Our latest result shows that $(\sqrt{2})^k < R(k,k)$. These are essentially the best known results on the size of $R(k,k)$, so there is a lot of progress to be made on Ramsey numbers.

Theorem 15.2 *Let n and m be two positive integers larger than 1, and let $m > 2.01 \log_2 n$. Then it is possible to color each edge of $K_{n,n}$ red or blue so that no $K_{m,m}$-subgraph with monochromatic edges is formed.*

Proof. The number of ways to 2-color the edges of a given $K_{m,m}$ subgraph of $K_{n,n}$ is 2^{m^2}, and two of these colorings result in monochromatic subgraphs. Therefore, the probability that at least one monochromatic $K_{m,m}$ is formed is at most $\binom{n}{m}^2 2^{1-m^2}$. Therefore, all we have to prove is

$$\binom{n}{m}^2 2^{1-m^2} < 1,$$

$$2\binom{n}{m}^2 < 2^{m^2}.$$

To see this, we insert two intermediate expressions as follows.

$$2\binom{n}{m}^2 < n^{2m} < (2^{m/2})^{2m} = 2^{m^2},$$

where the second inequality is a simple consequence of the relation between n and m. □

Note that the only property of the number 2.01 in the condition $m > 2.01 \log_2 n$ was that it is larger than 2. So any number $2+\epsilon$, with $\epsilon > 0$, would do.

Another way to formulate this same theorem is as follows. If $m > 2.01 \log_2 n$, then there exists a matrix of size $n \times n$ whose entries are either 0 or 1 having no $m \times m$ minor that consists of zeros only, or of ones only.

What is amazing about this result is that *nobody knows how to construct* that matrix, or *how* to color the edges of $K_{n,n}$ so that the requirements are fulfilled. In fact, the gap between what we *can* do and what we know is *possible* is rather large. The best construction known to this day for an $n \times n$ matrix with zeros and ones, and not having $m \times m$ homogeneous minors works for $m = c\sqrt{n}$, where c is a constant. This is much more than what we know is true, that is, $(2+\epsilon)\log_2 n$.

15.3 Independent Events

15.3.1 *The Notion of Independence and Bayes' Theorem*

Let us throw two dice at random. Let A be the event that the first die shows six, and let B be the event that the second die shows six. It is obvious that $P(A) = P(B) = 1/6$, and $P(A \cap B) = 1/36$. We see that $P(A) \cdot P(B) = P(A \cap B)$, and start wondering whether this is a coincidence. Now let us pick a positive integer from [12] at random. Let C be the event that this number is divisible by two, let D be the event that this number is divisible by three, and let F be the event that this number is divisible by four. Then we have $P(C) = 1/2$, $P(D) = 1/3$, and $P(F) = 1/4$. We also have $P(C \cap D) = 1/6$, and $P(D \cap F) = 1/12$, so the "product rule" seems to hold, but we also have $P(C \cap F) = P(F) = 1/4 \neq P(A)P(B)$, breaking the product rule.

Why is it that sometimes we find $P(A) \cdot P(B) = P(A \cap B)$, and sometimes we find $P(A) \cdot P(B) \neq P(A \cap B)$? As you have probably guessed, this is because sometimes the fact that A occurs make the occurrence of B more likely, or less likely, and sometimes does not alter the chance that B occurs at all. For example, if we choose an integer from 1 to 12, then the fact that it is divisible by two certainly makes it more likely that it is also divisible by four. Indeed, the number of all possible outcomes decreases from 12 to six, while that of favorable outcomes does not change. On the other hand, the fact that our number is divisible by two does not change its chances to be divisible by three. Indeed, the number of all outcomes decreases from 12 to six, but the number of favorable outcomes also decreases, from four to two.

This warrants the following two definitions.

Definition 15.2 If A and B are two events from the same sample space Ω, and $P(A \cap B) = P(A) \cdot P(B)$, then A and B are called *independent* events. Otherwise they are called *dependent*.

Definition 15.3 Let A and B be events from the same sample space, and assume $P(B) > 0$. Let

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Then $P(A|B)$ is called a *conditional probability*, and is read *"the probability of A given B"*.

That is, $P(A|B)$ is the probability of A given that B occurs. The following proposition is now immediate from the definitions.

Proposition 15.2 *The events A and B are independent if and only if $P(A|B) = P(A)$ holds.*

In other words, A and B are independent if and only if the occurrence of B does not make the occurrence of A any more likely, or any less likely.

Example 15.3 We toss a coin four times. We are not allowed to see the results, but we are told that there are at least two heads among the results. What is the probability that all four tosses resulted in heads?

Proof. Let A be the event that all four tosses are heads, and let B be the event that there are at least two heads. Then $A \cap B = A$, so $P(A|B) = P(A)/P(B)$. As the probability of getting a head at any one toss is 1/2, we have $P(A) = \frac{1}{2^4} = \frac{1}{16}$. There is 1/16 chance to get four heads, 4/16 chance to get three heads and one tail, and 6/16 chance to get two heads, two tails. Therefore, $P(B) = \frac{11}{2^4}$, and $P(A|B) = 1/11$. □

Example 15.4 Let $p = p_1p_2\cdots p_n$ be a randomly selected n-permutation. Let A be the event that $p_1 > p_2$, and let B be the event that $p_2 > p_3$. Compute $P(A|B)$, and decide if A and B are independent.

Proof. Clearly, $P(A) = P(B) = 1/2$ as can be seen by reversing the relevant pair of entries. On the other hand, $A \cap B$ is the event that $p_1 > p_2 > p_3$, which occurs in 1/6 of all permutations. Therefore,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3} \neq P(A),$$

so A and B are not independent. □

Your reaction to the previous example was probably something along the lines "Of course. If $p_1 > p_2$, then p_2 is smaller than normal, so it is less likely than normal that $p_2 > p_3$." While that argument works in this case, one should be extremely careful when injecting intuition into arguments involving conditional probabilities. The following example is a striking instance of this.

Example 15.5 A University has two colleges, the College of Liberal Arts, and the College of Engineering. Each college analyzed its own admission record and each college found that last year, a domestic applicant to the college had a larger chance to be admitted than an international applicant. Can we conclude that the same is true for the entire university? (Assume that applicants can only apply to one college.)

Proof. **No, we cannot.** A counterexample is shown in Figure 15.1. □

	Liberal Arts	Engineering	Entire University
Domestic applicants	Admitted: 10 Applied: 120 success rate:8.3%	Admitted: 10 Applied: 10 success rate: 100%	Admitted: 20 Applied: 130 success rate: 15.9%
International applicants	Admitted: 1 Applied: 15 success rate:6.7%	Admitted: 90 Applied: 100 success rate:90%	Admitted: 91 Applied: 115 success rate : 79.1%

Fig. 15.1 Not all that glitters is gold.

How is this very counterintuitive fact called *Simpson's paradox* possible? Some people do not believe it even when they see it with their own eyes. An imprecise, but conceptually correct, explanation is this. A much larger portion of the international applicants applied to Engineering, where the general rate of acceptance was higher. While it is true that the domestic students had an even higher acceptance rate in that college, it concerned only half of all domestic applicants, versus more than 85 percent of international applicants. In other words, more than 85 percent of all international applicants got into Engineering, whereas less than 16 percent of all domestic applicants did. This is a huge difference, and the College of Liberal Arts, with relatively few applicants, cannot make up for that.

In order to find a more precise explanation, we will need Bayes' Theorem.

Theorem 15.3 *[Bayes' Theorem] Let A and B be mutually exclusive events so that $A \cup B = \Omega$, and $P(A)P(B) > 0$ holds. Let C be any event. Then*

$$P(C) = P(C|A) \cdot P(A) + P(C|B) \cdot P(B). \tag{15.2}$$

In other words, the probability of C is the weighted average of its conditional probabilities, where the weights are the probabilities of the conditions.

Proof. As A and B are mutually exclusive, $A \cap C$ and $B \cap C$ are disjoint, and since $A \cup B = \Omega$, their union is exactly C. Therefore,

$$P(C) = P(C \cap A) + P(C \cap B),$$

and the proof follows as the first (resp. second) member of the right-hand side agrees with the first (resp. second) member of the right-hand side of 15.2. □

Now we are in a position to provide a deeper explanation for Example 15.5. Let A_1 (resp. B_1) be the event that an international (resp. domestic) applicant *applies* to the college of Liberal Arts, and define A_2 and B_2 similarly, for the college of Engineering. Let C_1 (resp. C_2) be the event that an international (resp. domestic) applicant is admitted to the *university*. Then Theorem 15.3 shows that

$$P(C_1) = P(C_1|A_1) \cdot P(A_1) + P(C_1|B_1) \cdot P(B_1),$$

and

$$P(C_2) = P(C_2|A_2) \cdot P(A_2) + P(C_2|B_2) \cdot P(B_2).$$

The criterion requiring that domestic students have larger chances to get accepted by any one college ensures that $P(C_1|A_1) < P(C_2|A_2)$, and $P(C_1|B_1) < P(C_2|B_2)$. It does not, however, say anything about $P(A_1)$ and $P(B_1)$. (We know that A_2 is the complement of A_1, and B_2 is the complement of B_1.) Therefore, we can choose A_1 and B_1 so that it is very advantageous for $P(C_1)$, and very bad for $P(C_2)$. We can do this by choosing $P(A_1)$ to be large if $P(C_1|A_1)$ is large, and by choosing $P(A_1)$ small if $P(C_1|A_1)$ is small. Similarly, we can choose $P(A_2)$ to be large if $P(C_1|A_1)$ is small, and vice versa.

In other words, weighted averages are a lot harder to control than unweighted averages. Indeed, if we impose the additional condition that $P(A_1) = P(B_1) = 1/2$, or even only the condition $P(A_1) = P(B_1)$, then the domestic students would have a greater chance to be admitted to the university.

15.3.2 *More Than Two Events*

It is not obvious at first sight how the independence of three or more events should be defined. We could require that $P(A_1 \cap A_2 \cdots A_n) = P(A_1) \cdot P(A_2) \cdot \cdots \cdot P(A_n)$. This, in itself, is not a very strong requirement, however. It holds whenever $P(A_1) = 0$, no matter how strongly the other variables depend on each other. To have some more local conditions, we can add the requirements that $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$. However, consider the following situation.

We select a positive integer from [10] at random. Let A be the event that this number is odd. Now let us select an integer from [20] at random, and let B be the event that this number is odd. Finally, let C be the event that the *difference* of the two selected integers is odd.

It is then clear that $P(A) = P(B) = P(C) = 1/2$, and also the events A, B, and C are *pairwise* independent, that is, any two of them are independent. However, $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C) = 1/8$. Therefore, we do not want to call these events independent, either.

We resolve these possible problems by requiring a very strong property for a set of events to be independent.

Definition 15.4 We say that the events $A_1, A_2, \cdots, A_n$ are independent if, for any nonempty subset $S = \{i_1, i_2, \cdots, i_k\} =\subseteq [n]$, we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \cdots \cdot P(A_{i_k}).$$

Theorem 15.3 is easy to generalize to more than two conditions.

Theorem 15.4 *[Bayes' Theorem, General Version] Let $A_1, A_2, \cdots, A_n$ be events in a sample space Ω so that $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$, and $A_i \cap A_j = \emptyset$ if $i \neq j$. Let $C \subset \Omega$ be any event. Then we have*

$$P(C) = \sum_{i=1}^{n} P(C|A_i)P(A_i).$$

Proof. Analogous to that of Theorem 15.3. □

15.4 Expected Values

A *random variable* is a *function* that is defined on a sample space Ω, and whose range is a set of numbers. For example, if Ω is the set of all graphs on n labeled vertices, we can define the random variable X by setting $X(G)$ to be the number of edges of G, or we can define the random variable Y by setting Y to be the number of connected components of G.

Just as for functions, we can define the sum and product of random variables over the same sample space the usual way, that is, $(X+Y)(u) = X(u) + Y(u)$, and $(X \cdot Y)(u) = X(u) \cdot Y(u)$.

Possibly the most important and useful parameter of a random variable is its *expected value*, or, in other words, *expectation*, or *average value*, or *mean*.

Definition 15.5 Let $X : \Omega \to \mathbf{R}$ be a random variable so that the set $S = \{X(u) | u \in \Omega\}$ is finite, that is, X only takes a finite number of values. Then the number

$$E(X) = \sum_{i \in S} i \cdot P(X = i)$$

is called the *expected value*, or *expectation* of X on Ω.

Here, and throughout this chapter, $P(X = i)$ is the probability of the event that $X(u) = i$. That is, $P(X = i) = \frac{|\{u \in \Omega | X(u) = i\}|}{|\Omega|}$.

In other words, $E(X)$ is the weighted average of all values X takes, with the weights being equal to the probability of X taking the corresponding value.

Remarks. Some probability variables can be defined over many different sample spaces. Our above example, the number of edges of a graph, can be defined not just over the space of all graphs on n vertices, or all *connected graphs* on n vertices, but also on all graphs on *at most* $3n$ vertices, and so on. In each case, the set $S = \{X(u) | u \in \Omega\}$ is different, therefore the expectation of X is also different. Therefore, if there is a danger of confusion, we write $E_\Omega(X)$, to denote where the expectation is taken. If there is no danger of confusion, however, we will only write $E(X)$, to alleviate notation.

Sometimes we announce both Ω and X in the same sentence as in "let $X(G)$ be the number of edges of a randomly selected connected graph G on n vertices." This means that Ω is the set of all connected graphs on n vertices, and $X(G)$ is the number of edges of the graph $G \in \Omega$.

It is possible to define the expectation of X in some cases when the set $S = \{X(u) | u \in \Omega\}$ is *not finite*. If S is a countably infinite set, we can define $E(X) = \sum_{i \in S} i \cdot P(X = i)$ as long as this infinite sum exists. See

Exercise 4 for an example. If S is not countable, the summation may be replaced by integration. Details can be found in any probability textbook.

Definition 15.6 The random variables X and Y are called *independent* if for all s and t, we have

$$P(X = s, Y = t) = P(X = s)P(Y = t).$$

15.4.1 *Linearity of Expectation*

For any real number c, we can define the random variable cX by setting $cX(u) = c(X(u))$ for all $u \in \Omega$. The following innocent-looking theorem proves to be extremely useful in enumerative combinatorics.

Theorem 15.5

(1) *Let X and Y be two random variables defined over the same space Ω. Then $E(X+Y) = E(X) + E(Y)$.*

(2) *Let X be a random variable, and let c be a real number. Then $E(cX) = cE(X)$.*

So "taking expectations" is a linear operator. The best feature of this theorem is that it does not require that X and Y be independent! No matter how deeply X and Y are intertwined, nor how hard it is to compute, say, the probability that $X = Y$, the expected value of $X + Y$ is always given by this simple formula.

Proof.

(1) Let $r \in \Omega$, then by definition we have $X(r)+Y(r) = (X+Y)(r)$, so $X(r)P(r) + Y(r)P(r) = (X+Y)(r)P(r)$. Adding these equations for all $r \in \Omega$, we get

$$\begin{aligned} E(X+Y) &= \sum_{r\in\Omega}(X+Y)(r)P(r) = \sum_{r\in\Omega} X(r)P(r) + \sum_{r\in\Omega} Y(r)P(r) \\ &= E(X) + E(Y). \end{aligned}$$

(2) Let $r \in \Omega$, then by definition we have $(cX)(r) = cX(r)$. Adding these equations for all $r \in \Omega$, we get

$$E(cX) = \sum_{r\in\Omega}(cX)(r)P(r) = c\sum_{r\in\Omega} X(r)P(r) = cE(X).$$

□

To see the surprising strength of Theorem 15.5, let $p = p_1p_2\cdots p_n$ be an n-permutation, and let us say that i is a valley if p_i is smaller than both of its neighbors, that is $p_i < p_{i-1}$, and $p_i < p_{i+1}$. We require $2 \leq i \leq n-1$ for i to be a valley.

Theorem 15.6 *Let $n \geq 2$ be a positive integer. Then on average, a randomly selected permutation of length n has $(n-2)/3$ valleys.*

Without Theorem 15.5, this would be a painful task. We would have to compute the number $v(j)$ of n-permutations with j valleys for each j, (a difficult task), then we would have to compute $\sum_j j \cdot \frac{v(j)}{n!}$. Theorem 15.5, however, turns the proof into a breeze.

Proof. Take $n-2$ different probability variables $Y_2, Y_3, \cdots, Y_{n-1}$, defined on the set of all n-permutations as follows. For an n-permutation p, let $Y_i(p) = 1$ if i is a valley, and let $Y_i(p) = 0$ otherwise. It is clear that for $2 \leq i \leq n-1$, every p_i has a $1/3$ chance to be the smallest of the set $\{p_{i-1}, p_i, p_{i+1}\}$. Therefore,

$$E(Y_i) = \frac{1}{3}\cdot 1 + \frac{2}{3}\cdot 0 = \frac{1}{3}.$$

Define $Y = Y_2 + Y_3 + \cdots + Y_{n-1}$. It is clear that $Y(p)$ is the number of valleys of p. Then Theorem 15.5 implies

$$E(Y) = \sum_{i=2}^{n-1} E(Y_i) = (n-2)\cdot E(Y_1) = \frac{n-2}{3}.$$

□

Variables similar to Y_i, that is, variables that take value 1 if a certain event occurs, and value 0 otherwise, are called *indicator variables*.

Theorem 15.7 *The expected value of the number of fixed points in a randomly selected n-permutation is 1.*

Proof. We define n different probability variables $X_1, X_2, \cdots, X_n$ on the set of all n-permutations as follows. For an n-permutation p, let $X_i(p) = 1$ if $p_i = i$, that is, when p has a fixed point at position i, and let $X_i(p) = 0$ otherwise.

As p_i is equally likely to take any value $j \in [n]$, it has a $1/n$ chance to be equal to i. Therefore,

$$E(X_i) = \frac{1}{n} \cdot 1 + \frac{n-1}{n} \cdot 0 = \frac{1}{n},$$

for all $i \in [n]$. Now define $X = X_1 + X_2 + \cdots + X_n$; it is then clear that $X(p)$ is precisely the number of fixed points of p. On the other hand, applying Theorem 15.5, we get

$$E(X) = \sum_{i=1}^{n} E(X_i) = n \cdot E(X_1) = n \cdot \frac{1}{n} = 1, \qquad (15.3)$$

which was to be proved. □

15.4.2 *Existence Proofs Using Expectation*

It is common sense knowledge that the average of a set of numbers is never larger than the largest of those numbers. This is true for weighted averages as well.

Theorem 15.8 *Let $X : \Omega \to \mathbf{R}$ be a random variable so that the set $S = \{X(u) | u \in \Omega\}$ is finite, and let j be the largest element of S. Then we have*

$$j \geq E(X).$$

Proof. Using the definition of $E(X)$, we have

$$E(X) = \sum_{i \in S} i \cdot P(X = i) \leq j \sum_{i \in S} P(X = i) = j.$$ □

We show two applications of this idea. The first shows that a simple graph will always contain a large bipartite subgraph.

Theorem 15.9 *Let G be a simple graph with vertex set $[n]$, and m edges. Then G contains a bipartite subgraph with more than $m/2$ edges.*

Proof. Let us split the vertices of G into two disjoint nonempty subsets A and B. Then A and B span a bipartite subgraph H of G. (We omit the edges within A and within B.) Let Ω be the set of $2^{n-1} - 1$ different bipartite subgraphs we get this way. Let $X(H)$ be the number of edges in H.

On the other hand, let us number the edges of G from one through m, and let $X_i = 1$ if the edge i has one vertex in A, and one in B, and let $X_i = 0$ otherwise.

What is $P(X_i = 1)$? To get such a subdivision of $[n]$, first we put the two endpoints of the edge i to different subsets, then split the remaining $n-2$-element vertex set in any of 2^{n-2} ways. Therefore, $P(X_i = 1) = \frac{2^{n-2}}{2^{n-1}-1}$, and $P(X_i = 0) = \frac{2^{n-2}-1}{2^{n-1}-1}$. This implies

$$E(X_i) = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = \frac{2^{n-2}}{2^{n-1}-1} > \frac{1}{2}.$$

We can repeat this argument for all edges. Then we note that $X = X_1 + X_2 + \cdots + X_m$, so Theorem 15.5 implies

$$E(X) = \sum_{i=1}^{m} E(X_i) = m \cdot E(X_1) > \frac{m}{2}.$$

As the expected value of the number of edges in these bipartite subgraphs of G is more than $m/2$, it follows from Theorem 15.8 that there is at least one bipartite subgraph of G with more than $m/2$ edges. □

The next example is related to a well-known problem in complexity theory, the so-called "Betweenness Problem".

Example 15.6 We are given a list $L = (L_1, L_2, \cdots, L_k)$ of ordered triples $L_i = (a_i, b_i, c_i)$, so that for any i, the numbers a_i, b_i, and c_i are distinct elements of $[n]$. It is possible, however, that symbols with *different* indices i and j denote the same number.

Let $p = p_1 p_2 \cdots p_n$ be an n-permutation. We say that p satisfies L_i if the entry b_i is *between* a_i and c_i in p. (It does not matter whether the order of these three entries in p is $a_i b_i c_i$ or $c_i b_i a_i$.)

Prove that there exists an n-permutation p that satisfies at least one third of all L_i in any given list L.

Proof. Let Y_i be the indicator variable of the event that a randomly chosen n-permutation satisfies L_i. Then clearly, $P(Y_i = 1) = \frac{1}{3}$ as each of a_i, b_i and c_i has the same chance to be in the middle. Therefore, $E(Y_i) = \frac{1}{3}$.

Now if $Y = \sum_{i=1}^{k} Y_i$, then Y is the number of L_i in L that are satisfied by p. Theorem 15.5 then implies

$$E(Y) = \sum_{i=1}^{k} E(Y_i) = \frac{k}{3},$$

and our claim follows from Theorem 15.8. Let Y_i be the indicator variable of the event that a randomly chosen n-permutation satisfies L_i. Then clearly, $P(Y_i = 1) = \frac{1}{3}$ as each of a_i, b_i and c_i has the same chance to be in the middle. Therefore, $E(Y_i) = \frac{1}{3}$. Now let $Y = \sum_{i=1}^{k} Y_i$, then Y is the number of L_i in L that are satisfied by p. Theorem 15.5 then implies

$$E(Y) = \sum_{i=1}^{k} E(Y_i) = \frac{k}{3},$$

and our claim follows from Theorem 15.8. □

15.4.3 *Conditional Expectation*

Another way of computing the expectation of a variable is by using conditional expectations. $E(X|A)$ is the expected value of X given that event A occurs. Accordingly, $E(X|A)$ is defined by replacing the absolute probabilities in the definition of $E(X)$ by probabilities conditional on the occurrence of A. In other words,

$$E(X|A) = \sum_i i \cdot P(X = i|A),$$

where i ranges through all values X takes with a positive probability, given that A occurs.

We can then extend Theorem 15.4 to expectations as follows.

Theorem 15.10 *Let X be a random variable, and let $A_1, A_2, \cdots, A_n$ be events in a sample space Ω so that $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$, and $A_i \cap A_j = \emptyset$ if $i \neq j$. Then we have*

$$E(X) = \sum_{i=1}^{n} E(X|A_i)P(A_i).$$

Proof. This follows immediately from Theorem 15.4. Just let C be the event $X = j$ in that theorem. Then multiply both sides by j, and sum over all values of j taken by X with a positive probability. □

Example 15.7 We throw a die three times. Provided that the first throw was at least four, what is the expectation of the number of times a throw resulted in an even number?

Proof. If the first throw was an even number, then the expected value of even results is obviously two as it is one on the last two throws. If the first throw was an odd number, then this expectation is 1. Therefore, Theorem 15.10 implies

$$E(X) = \sum_{i=1}^{2} E(X|A_i)P(A_i) = \frac{2}{3} \cdot 2 + \frac{1}{3} \cdot 1 = \frac{5}{3}.$$

□

In this problem, it was very easy to compute the probabilities $P(A_i)$. The following problem is a little bit less obvious in that aspect.

Example 15.8 Our football team wins each game with 3/4 probability. What is our expected value of wins in a 12-game season if we know that we won at least three of the first four games?

Proof. We either won three games (event A_1), or four games (event A_2) out of the first four games. If we disregard the condition that we won at least three games out of the first four (event B), we have $P(A_1) = 4 \cdot \frac{1}{4}(\frac{3}{4})^3 == \frac{27}{64}$, and $P(A_2) = (\frac{3}{4})^4 = \frac{81}{256}$. That condition, however, leads us to the conditional probabilities

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{\frac{27}{64}}{\frac{27}{64} + \frac{81}{256}} = \frac{4}{7},$$

and

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{3}{7}.$$

In this problem we assume that B occurred, that is, B is our sample space. To emphasize this, we will write $P_B(A_i)$ instead of $P(A_i|B)$. We denote the expectations accordingly.

In the last eight games of the season, the expected number of our wins is certainly $8 \cdot \frac{3}{4} = 6$, by Theorem 15.5. Therefore, denoting the number of our wins by X, Theorem 15.10 shows

$$E_B(X) = E_B(X|A_1)P_B(A_1) + E_B(X|A_2)P_B(A_2) = 9 \cdot \frac{4}{7} + 10 \cdot \frac{3}{7} = 9\frac{3}{7}. \square$$

We see that this expectation is larger than 9, the expectation without the condition that we won at least three of the first four games. This is because that condition allowed us to win all four of those games, which is better than our general performance.

Notes

This Chapter was not as much on Probability Theory itself as on the applications of Probability in Combinatorics. While there are plenty of textbooks on Probability Theory itself, there are not as many on Discrete Probability, that is, when Ω is finite. A very accessible introductory book in that field is [9]. As far as the Probabilistic Method in Combinatorics goes, a classic is [1].

Exercises

(1) Let p_n be the probability that a random text of n letters has a substring of consecutive letters that reads "Probability is fun". Prove that $\lim_{n\to\infty} = 1$.

(2) A big corporation has four level of commands. The CEO is at the top, (level 1) she has some direct subordinates (level 2), who in turn have their own direct subordinates (level 3), and even those people have their own direct subordinates (level 4). Nobody, however, has more direct subordinates than his immediate supervisor. Is it true that the average number of direct subordinates of an officer on level i is always higher than the average number of direct subordinates of an officer on level $i+1$?

(3) A women's health clinic has four doctors, and each patient is assigned to one of them. If a patient gives birth between 8am and 4pm, then her chance of being attended by her assigned doctor is 3/4, otherwise it is 1/4. What is the probability that a patient is attended by her assigned doctor when she gives birth?

(4) We toss a coin a finite number of times. Let S denote the sequence of results. Set $X(S) = i$ if a head occurs in position i first. Find $E_\Omega(X)$, where Ω is the set of all finite outcome sequences.

(5) Show that for any n, there exist n events so that any $n-1$ of them are independent, but the n events are not.

(6) At a certain university, a randomly selected student who has just enrolled has 66 percent chance to graduate in four years, but if he successfully completes all freshmen courses in his first year, then this chance goes up to 90 percent. Among those failing to complete at least one freshmen course in their first year, the 4-year-graduation rate is 50 percent. What is the percentage of all students who cannot complete all freshmen courses in their first year?

(7) We select an element of $[100]$ at random. Let A be the event that this integer is divisible by 3, and let B be the event that this integer is divisible by 7. Are A and B independent?

(8) Six football teams participate in a round robin tournament. Any two teams play each other exactly once. We say that three teams *beat each other* if in their games played against each other, each team got one victory and one loss. What is the expected number of triples of teams who beat each other? Assume that each game is a toss-up, that is, each team has 50 percent chance to win any of its games.

(9) Solve the previous exercise if one of the teams is so good that it wins its games by 90 percent probability.

(10) What is the expected value of the number of digits equal to 3 in a 4-digit positive integer?

(11) Let $X(\alpha)$ be the first part of a randomly selected weak composition α of n. Find $E(X)$.

(12) Let $Y(\alpha)$ be the number of parts in a randomly selected weak composition α of n. Find $E(Y)$.

(13) Let π be a randomly selected partition of the integer n. Let $X(p)$ be the first part of π, and let $Y(p)$ be the number of parts in π. Find $E(X) - E(Y)$.

(14) Let $p = p_1p_2\cdots p_n$ be an n-permutation. The index i is called an *excedance* of p if $p(i) > i$. How many excedances does the average

n-permutation have?

(15) Let k be any positive integer, and let $n \geq k$. Let Y be the number of k-cycles in a randomly selected n-permutation. Find $E(Y)$.

(16) Recall from Chapter 14 that $S_n(1234) < S_n(1324)$ if $n \geq 7$. Let n be a fixed integer so that $n \geq 7$. Let A be the event that an n-permutation contains a 1234-pattern, and let B be the event that an n-permutation contains a 1324-pattern. Similarly, let X, (resp. Y) be the number of 1234-patterns (resp. 1324-patterns) in a randomly selected n-permutation. What is larger, $E(X|A)$ or $E(Y|B)$?

(17) Prove that there is a tournament on n vertices that contains at least $\frac{n!}{2^{n-1}}$ Hamiltonian paths. What can we say about the number of Hamiltonian cycles?

(18) Let Y be a probability variable. Then $Var(Y) = E((Y - E(Y))^2)$ is called the *variance* of Y.

a. Prove that $Var(Y) = E(Y^2) - E(Y)^2$.

b. Let $X(p)$ be the number of fixed points of a randomly selected n-permutation p. Prove that $Var(X) = 1$.

(19) For $i \in [n]$, define X_i as in the proof of Theorem 15.7. Are the X_i independent?

(20) Let X and Y be two independent random variables defined on the same space. Prove that $Var(X + Y) = Var(X) + Var(Y)$.

(21) We are given a list $L = (L_1, L_2, \cdots, L_k)$ of ordered 4-tuples $L_i = (a_i, b_i, c_i, d_i)$, so that for any i, the numbers a_i, b_i, c_i, and d_i are distinct elements of $[n]$. It is possible, however, that symbols with *different* indices i and j denote the same number.
Let $p = p_1p_2 \cdots p_n$ be an n-permutation. We say that p satisfies L_i if the substring of p that stretches from a_i to b_i *does not intersect* the substring of p that stretches from c_i to d_i. (It could be that a_i is on the right of b_i, or c_i is on the right of d_i.)
Prove that there exists an n-permutation p that satisfies at least one third of all L_i in any given list L.

Supplementary Exercises

(22) Prove that it is possible to 2-color the integers from 1 to 1000 so that no monochromatic arithmetic progression of length 17 is formed.

(23) Is it true that if the occurrence of A makes B more likely to occur, then the occurrence of B also makes A more likely to occur?

(24) Let S be an $n \times n$ magic square (see Exercise 24 in Chapter 3) with line sum r. Let A be the event that each entry of the first row is at least $\frac{r}{2n}$, and let B be the event that each element of the second row is at least $\frac{r}{2n}$. Is the following argument correct?
"We have $P(B|A) < P(B)$. Indeed, if A occurs, then the entries of the first row are all larger than normal, so each entry of the second row must be smaller than normal, because the sum of each column is fixed."

(25) Can two events be at the same time mutually exclusive and independent?

(26) Adam and Brandi are playing the following game. They write each integer from 1 through 100 on a piece a paper, then they randomly select a piece of paper, and then another one. They add the two integers that are written on the two pieces of paper, and if the sum is even, then Adam wins, if not, then Brandi. Is this a fair game?

(27) Replace 100 by n in the previous exercise. For which positive integers n will the game be fair?

(28) A dealership has n cars. An employee with a sense of humor takes all n keys, puts one of them in each car at random, then locks the doors of all cars. When the owner of the dealership discovers the problem, he calls a locksmith. He tells him to break into a car, then use the key found in that car to open another, and so on. If and when the keys already recovered by this procedure cannot open any new cars, the locksmith is to break into another car. This algorithm goes on until all cars are open.

a. What is the probability that the locksmith will only have to break into one car?

b. + What is the probability that the locksmith will have to break into two cars only?

c. + What is the probability that the locksmith will have to break into at most k cars?

(29) There are 16 disks in a box. Five of them are painted red, five of them are painted blue, and six are painted red on one side, and blue on the other side. We are given a disk at random, and see that one of its sides is red. Is the other side of this disk more likely to be red or blue?

(30) There are ten disks in a basket, two of them are blue on both sides, three of them are red on both sides, and the remaining five are red on one side, and blue on the other side. One disk is drawn at random, and we have to guess the color of its back. Does it help if we know the color of its front?

(31) A pack of cards consists of 100 cards, two of them are black kings. We shuffle the cards, then we start dealing them until we draw a black king. Which is the step where this is most likely to occur?

(32) Let $p = p_1 p_2 \cdots p_n$ be an n-permutation. We say that p get changes direction at position i, if either $p_{i-1} < p_i > p_{i+1}$, or $p_{i-1} > p_i < p_{i+1}$, in other words, when p_i is either a *peak* or a *valley*. We say that p has k *runs* if there are $k-1$ indices i so that p changes direction at these positions. For example, $p = 3561247$ has 3 runs as p changes direction when $i = 3$ and when $i = 4$.
What is the average number of runs in a randomly selected n-permutation?

Solutions to Exercises

(1) First, we note that the sequence $\{p_n\}$ is increasing. Indeed, $p_{n+1} = p_n + q_n$, where q_n is the probability of the event that the set of the first n letters does not contain the required sentence, but that of the first $n+1$ letters does.
It is therefore sufficient to show that the sequence $\{p_n\}$ has a subsequence that converges to 1. Such a subsequence is $r_n = p_{16n}$. (Note that the sentence "Probability is fun" contains 16 letters.)
Let a be the probability of the event that a randomly selected 16-letter string is not our required sentence. Then $a < 1$. On the other hand, $r_n \geq 1 - a^n$ as we can split a $16n$-letter string into n strings of length 16, each of which has a chance to be something else than

our sentence. So we have

$$1 - a^n \leq r_n \leq 1,$$

and our claim follows by the squeeze principle as $a^n \to 0$.

(2) That is not true. Figure 15.2 shows a counterexample. Indeed, the average number of direct subordinates of level-2 officers is $6/4 = 1.5$, while that of level-3 officers is $10/6 = 1.66$.

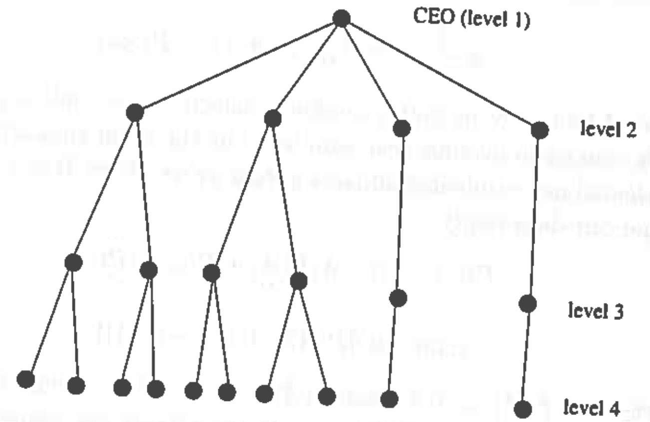

Fig. 15.2 A counterexample for Exercise 2.

(3) There is 1/3 chance that a given patient gives birth between 8am and 4pm, and there is 2/3 chance that she gives birth between 4pm and 8am. Therefore, Bayes' theorem shows that the answer is $\frac{1}{3} \cdot \frac{3}{4} + \frac{2}{3} \cdot \frac{1}{4} = \frac{5}{12}$.

(4) The only way for the first head to occur in position i is to have a tail in each of the first $i-1$ positions, then a head in position i. The chance of this happening is $1/2^i$. Therefore, we have

$$E(X) = \sum_{i=1}^{\infty} \frac{i}{2^i} = 2.$$

We used the fact that $\sum_{n \geq 1} nx^n = \frac{x}{(1-x)^2}$. This has been proved in two different ways in Exercise 25 of Chapter 4.

(5) Let us throw a die $n-1$ times, and for $1 \leq i \leq n-1$, denote A_i the event that throw i results in an even number. Finally, let A_n