

Discrete Probability

- 6.1 An Introduction to Discrete Probability
- 6.2 Probability Theory
- 6.3 Bayes' Theorem
- 6.4 Expected Value and Variance

Combinatorics and probability theory share common origins. The theory of probability was first developed more than 300 years ago, when certain gambling games were analyzed. Although probability theory was originally invented to study gambling, it now plays an essential role in a wide variety of disciplines. For example, probability theory is extensively applied in the study of genetics, where it can be used to help understand the inheritance of traits. Of course, probability still remains an extremely popular part of mathematics because of its applicability to gambling, which continues to be an extremely popular human endeavor.

In computer science, probability theory plays an important role in the study of the complexity of algorithms. In particular, ideas and techniques from probability theory are used to determine the average-case complexity of algorithms. Probabilistic algorithms can be used to solve many problems that cannot be easily or practically solved by deterministic algorithms. In a probabilistic algorithm, instead of always following the same steps when given the same input, as a deterministic algorithm does, the algorithm makes one or more random choices, which may lead to different output. In combinatorics, probability theory can even be used to show that objects with certain properties exist. The probabilistic method, a technique in combinatorics introduced by Paul Erdős and Alfréd Rényi, shows that an object with a specified property exists by showing that there is a positive probability that a randomly constructed object has this property. Probability theory can help us answer questions that involve uncertainty, such as determining whether we should reject an incoming mail message as spam based on the words that appear in the message.

6.1 An Introduction to Discrete Probability

Introduction

Probability theory dates back to the seventeenth century when the French mathematician Blaise Pascal determined the odds of winning some popular bets based on the outcome when a pair of dice is repeatedly rolled. In the eighteenth century, the French mathematician Laplace, who also studied gambling, defined the probability of an event as the number of successful outcomes divided by the number of possible outcomes. For instance, the probability that a die comes up an odd number when it is rolled is the number of successful outcomes—namely, the number of ways it can come up odd—divided by the number of possible outcomes—namely, the number of different ways the die can come up. There are a total of six possible outcomes—namely, 1, 2, 3, 4, 5, and 6—and exactly three of these are successful outcomes—namely, 1, 3, and 5. Hence, the probability that the die comes up an odd number is $3/6 = 1/2$. (Note that it has been assumed that all possible outcomes are equally likely, or, in other words, that the die is fair.)

In this section we will restrict ourselves to experiments that have finitely many, equally likely, outcomes. This permits us to use Laplace's definition of the probability of an event. We will continue our study of probability in Section 6.2, where we will study experiments with finitely many outcomes that are not necessarily equally likely. In Section 6.2 we will also introduce some key concepts in probability theory, including conditional probability, independence of events, and random variables. In Section 6.4 we will introduce the concepts of the expectation and variance of a random variable.

Finite Probability

An **experiment** is a procedure that yields one of a given set of possible outcomes. The **sample space** of the experiment is the set of possible outcomes. An **event** is a subset of the sample space. Laplace's definition of the probability of an event with finitely many possible outcomes will now be stated.

DEFINITION 1 If S is a finite sample space of equally likely outcomes, and E is an event, that is, a subset of S , then the *probability* of E is $p(E) = \frac{|E|}{|S|}$.

Examples 1–8 illustrate how the probability of an event is found.

EXAMPLE 1 An urn contains four blue balls and five red balls. What is the probability that a ball chosen from the urn is blue?

Extra Examples 

Solution: To calculate the probability, note that there are nine possible outcomes, and four of these possible outcomes produce a blue ball. Hence, the probability that a blue ball is chosen is $4/9$. ◀

EXAMPLE 2 What is the probability that when two dice are rolled, the sum of the numbers on the two dice is 7?

Solution: There are a total of 36 equally likely possible outcomes when two dice are rolled. (The product rule can be used to see this; because each die has six possible outcomes, the total number of outcomes when two dice are rolled is $6^2 = 36$.) There are six successful outcomes, namely, (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1), where the values of the first and second dice are represented by an ordered pair. Hence, the probability that a seven comes up when two fair dice are rolled is $6/36 = 1/6$. ◀

Links 

Lotteries are extremely popular. We can easily compute the odds of winning different types of lotteries.

EXAMPLE 3 In a lottery, players win a large prize when they pick four digits that match, in the correct order, four digits selected by a random mechanical process. A smaller prize is won if only three digits

Links 

PIERRE-SIMON LAPLACE (1749–1827) Pierre-Simon Laplace came from humble origins in Normandy. In his childhood he was educated in a school run by the Benedictines. At 16 he entered the University of Caen intending to study theology. However, he soon realized his true interests were in mathematics. After completing his studies, he was named a provisional professor at Caen, and in 1769 he became professor of mathematics at the Paris Military School.

Laplace is best known for his contributions to celestial mechanics, the study of the motions of heavenly bodies. His *Traité du Mécanique Céleste* is considered one of the greatest scientific works of the early nineteenth century. Laplace was one of the founders of probability theory and made many contributions to mathematical statistics. His work in this area is documented in his book *Théorie Analytique des Probabilités*, in which he defined the probability of an event as the ratio of the number of favorable outcomes to the total number of outcomes of an experiment.

Laplace was famous for his political flexibility. He was loyal, in succession, to the French Republic, Napoleon, and King Louis XVIII. This flexibility permitted him to be productive before, during, and after the French Revolution.

are matched. What is the probability that a player wins the large prize? What is the probability that a player wins the small prize?

Solution: There is only one way to choose all four digits correctly. By the product rule, there are $10^4 = 10,000$ ways to choose four digits. Hence, the probability that a player wins the large prize is $1/10,000 = 0.0001$.

Players win the smaller prize when they correctly choose exactly three of the four digits. Exactly one digit must be wrong to get three digits correct, but not all four correct. By the sum rule, to find the number of ways to choose exactly three digits correctly we add the number of ways to choose four digits matching the digits picked in all but the i th position, for $i = 1, 2, 3, 4$.

To count the number of successes with the first digit incorrect, note that there are nine possible choices for the first digit (all but the one correct digit), and one choice for each of the other digits, namely, the correct digits for these slots. Hence, there are nine ways to choose four digits where the first digit is incorrect, but the last three are correct. Similarly, there are nine ways to choose four digits where the second digit is incorrect, nine with the third digit incorrect, and nine with the fourth digit incorrect. Hence, there is a total of 36 ways to choose four digits with exactly three of the four digits correct. Thus, the probability that a player wins the smaller prize is $36/10,000 = 9/2500 = 0.0036$. ◀

EXAMPLE 4 There are many lotteries now that award enormous prizes to people who correctly choose a set of six numbers out of the first n positive integers, where n is usually between 30 and 60. What is the probability that a person picks the correct six numbers out of 40?

Solution: There is only one winning combination. The total number of ways to choose six numbers out of 40 is

$$C(40, 6) = \frac{40!}{34! 6!} = 3,838,380.$$

Consequently, the probability of picking a winning combination is $1/3,838,380 \approx 0.00000026$. (Here the symbol \approx means approximately equal to.) ◀



Poker, and other card games, are growing in popularity. To win at these games it helps to know the probability of different hands. We can find the probability of specific hands that arise in card games using the techniques developed so far. A deck of cards contains 52 cards. There are 13 different kinds of cards, with four cards of each kind. (Among the terms commonly used instead of “kind” are “rank,” “face value,” “denomination,” and “value.”) These kinds are twos, threes, fours, fives, sixes, sevens, eights, nines, tens, jacks, queens, kings, and aces. There are also four suits: spades, clubs, hearts, and diamonds, each containing 13 cards, with one card of each kind in a suit. In many poker games, a hand consists of five cards.

EXAMPLE 5 Find the probability that a hand of five cards in poker contains four cards of one kind.

Solution: By the product rule, the number of hands of five cards with four cards of one kind is the product of the number of ways to pick one kind, the number of ways to pick the four of this kind out of the four in the deck of this kind, and the number of ways to pick the fifth card. This is

$$C(13, 1)C(4, 4)C(48, 1).$$

By Example 11 in Section 5.3 there are $C(52, 5)$ different hands of five cards. Hence, the probability that a hand contains four cards of one kind is

$$\frac{C(13, 1)C(4, 4)C(48, 1)}{C(52, 5)} = \frac{13 \cdot 1 \cdot 48}{2,598,960} \approx 0.00024. \quad \blacktriangleleft$$

EXAMPLE 6 What is the probability that a poker hand contains a full house, that is, three of one kind and two of another kind?

Solution: By the product rule, the number of hands containing a full house is the product of the number of ways to pick two kinds in order, the number of ways to pick three out of four for the first kind, and the number of ways to pick two out of four for the second kind. (Note that the order of the two kinds matters, because, for instance, three queens and two aces is different than three aces and two queens.) We see that the number of hands containing a full house is

$$P(13, 2)C(4, 3)C(4, 2) = 13 \cdot 12 \cdot 4 \cdot 6 = 3744.$$

Because there are 2,598,960 poker hands, the probability of a full house is

$$\frac{3744}{2,598,960} \approx 0.0014. \quad \blacktriangleleft$$

EXAMPLE 7 What is the probability that the numbers 11, 4, 17, 39, and 23 are drawn in that order from a bin containing 50 balls labeled with the numbers 1, 2, ..., 50 if (a) the ball selected is not returned to the bin before the next ball is selected and (b) the ball selected is returned to the bin before the next ball is selected?

Solution: (a) By the product rule, there are $50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 = 254,251,200$ ways to select the balls because each time a ball is drawn there is one fewer ball to choose from. Consequently, the probability that 11, 4, 17, 39, and 23 are drawn in that order is $1/254,251,200$. This is an example of **sampling without replacement**.

(b) By the product rule, there are $50^5 = 312,500,000$ ways to select the balls because there are 50 possible balls to choose from each time a ball is drawn. Consequently, the probability that 11, 4, 17, 39, and 23 are drawn in that order is $1/312,500,000$. This is an example of **sampling with replacement**. \blacktriangleleft

The Probability of Combinations of Events

We can use counting techniques to find the probability of events derived from other events.

THEOREM 1 Let E be an event in a sample space S . The probability of the event \overline{E} , the complementary event of E , is given by

$$p(\overline{E}) = 1 - p(E).$$

Proof: To find the probability of the event \overline{E} , note that $|\overline{E}| = |S| - |E|$. Hence,

$$p(\overline{E}) = \frac{|S| - |E|}{|S|} = 1 - \frac{|E|}{|S|} = 1 - p(E). \quad \blacktriangleleft$$

There is an alternative strategy for finding the probability of an event when a direct approach does not work well. Instead of determining the probability of the event, the probability of its complement can be found. This is often easier to do, as Example 8 shows.

EXAMPLE 8 A sequence of 10 bits is randomly generated. What is the probability that at least one of these bits is 0?

Solution: Let E be the event that at least one of the 10 bits is 0. Then \overline{E} is the event that all the bits are 1s. Because the sample space S is the set of all bit strings of length 10, it follows that

$$\begin{aligned} p(E) &= 1 - p(\overline{E}) = 1 - \frac{|\overline{E}|}{|S|} = 1 - \frac{1}{2^{10}} \\ &= 1 - \frac{1}{1024} = \frac{1023}{1024}. \end{aligned}$$

Hence, the probability that the bit string will contain at least one 0 bit is $1023/1024$. It is quite difficult to find this probability directly without using Theorem 1. ◀

We can also find the probability of the union of two events.

THEOREM 2 Let E_1 and E_2 be events in the sample space S . Then

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2).$$

Proof: Using the formula given in Section 2.2 for the number of elements in the union of two sets, it follows that

$$|E_1 \cup E_2| = |E_1| + |E_2| - |E_1 \cap E_2|.$$

Hence,

$$\begin{aligned} p(E_1 \cup E_2) &= \frac{|E_1 \cup E_2|}{|S|} \\ &= \frac{|E_1| + |E_2| - |E_1 \cap E_2|}{|S|} \\ &= \frac{|E_1|}{|S|} + \frac{|E_2|}{|S|} - \frac{|E_1 \cap E_2|}{|S|} \\ &= p(E_1) + p(E_2) - p(E_1 \cap E_2). \end{aligned}$$

◀

EXAMPLE 9 What is the probability that a positive integer selected at random from the set of positive integers not exceeding 100 is divisible by either 2 or 5?



Solution: Let E_1 be the event that the integer selected is divisible by 2, and let E_2 be the event that it is divisible by 5. Then $E_1 \cup E_2$ is the event that it is divisible by either 2 or 5. Also,

$E_1 \cap E_2$ is the event that it is divisible by both 2 and 5, or equivalently, that it is divisible by 10. Because $|E_1| = 50$, $|E_2| = 20$, and $|E_1 \cap E_2| = 10$, it follows that

$$\begin{aligned} p(E_1 \cup E_2) &= p(E_1) + p(E_2) - p(E_1 \cap E_2) \\ &= \frac{50}{100} + \frac{20}{100} - \frac{10}{100} = \frac{3}{5}. \end{aligned}$$

Probabilistic Reasoning

A common problem is determining which of two events is more likely. Analyzing the probabilities of such events can be tricky. Example 10 describes a problem of this type. It discusses a famous problem originating with the television game show *Let's Make a Deal*.

EXAMPLE 10 **The Monty Hall Three-Door Puzzle** Suppose you are a game show contestant. You have a chance to win a large prize. You are asked to select one of three doors to open; the large prize is behind one of the three doors and the other two doors are losers. Once you select a door, the game show host, who knows what is behind each door, does the following. First, whether or not you selected the winning door, he opens one of the other two doors that he knows is a losing door (selecting at random if both are losing doors). Then he asks you whether you would like to switch doors. Which strategy should you use? Should you change doors or keep your original selection, or does it not matter?



Solution: The probability you select the correct door (before the host opens a door and asks you whether you want to change) is $1/3$, because the three doors are equally likely to be the correct door. The probability this is the correct door does not change once the game show host opens one of the other doors, because he will always open a door that the prize is not behind.

The probability that you selected incorrectly is the probability the prize is behind one of the two doors you did not select. Consequently, the probability that you selected incorrectly is $2/3$. If you selected incorrectly, when the game show host opens a door to show you that the prize is not behind it, the prize is behind the other door. You will always win if your initial choice was incorrect and you change doors. So, by changing doors, the probability you win is $2/3$. In other words, you should always change doors when given the chance to do so by the game show host. This doubles the probability that you will win. (A more rigorous treatment of this puzzle can be found in Exercise 15 of Section 6.3.)

Exercises

- What is the probability that a card selected from a deck is an ace?
- What is the probability that a die comes up six when it is rolled?
- What is the probability that a randomly selected integer chosen from the first 100 positive integers is odd?
- What is the probability that a randomly selected day of the year (from the 366 possible days) is in April?
- What is the probability that the sum of the numbers on two dice is even when they are rolled?
- What is the probability that a card selected from a deck is an ace or a heart?
- What is the probability that when a coin is flipped six times in a row, it lands heads up every time?
- What is the probability that a five-card poker hand contains the ace of hearts?
- What is the probability that a five-card poker hand does not contain the queen of hearts?
- What is the probability that a five-card poker hand contains the two of diamonds and the three of spades?
- What is the probability that a five-card poker hand contains the two of diamonds, the three of spades, the six of hearts, the ten of clubs, and the king of hearts?

12. What is the probability that a five-card poker hand contains exactly one ace?
13. What is the probability that a five-card poker hand contains at least one ace?
14. What is the probability that a five-card poker hand contains cards of five different kinds?
15. What is the probability that a five-card poker hand contains two pairs (that is, two of each of two different kinds and a fifth card of a third kind)?
16. What is the probability that a five-card poker hand contains a flush, that is, five cards of the same suit?
17. What is the probability that a five-card poker hand contains a straight, that is, five cards that have consecutive kinds? (Note that an ace can be considered either the lowest card of an A-2-3-4-5 straight or the highest card of a 10-J-Q-K-A straight.)
18. What is the probability that a five-card poker hand contains a straight flush, that is, five cards of the same suit of consecutive kinds?
- *19. What is the probability that a five-card poker hand contains cards of five different kinds and does not contain a flush or a straight?
20. What is the probability that a five-card poker hand contains a royal flush, that is, the 10, jack, queen, king, and ace of one suit?
21. What is the probability that a die never comes up an even number when it is rolled six times?
22. What is the probability that a positive integer not exceeding 100 selected at random is divisible by 3?
23. What is the probability that a positive integer not exceeding 100 selected at random is divisible by 5 or 7?
24. Find the probability of winning the lottery by selecting the correct six integers, where the order in which these integers are selected does not matter, from the positive integers not exceeding
 - a) 30. b) 36. c) 42. d) 48.
25. Find the probability of winning the lottery by selecting the correct six integers, where the order in which these integers are selected does not matter, from the positive integers not exceeding
 - a) 50. b) 52. c) 56. d) 60.
26. Find the probability of selecting none of the correct six integers, where the order in which these integers are selected does not matter, from the positive integers not exceeding
 - a) 40. b) 48. c) 56. d) 64.
27. Find the probability of selecting exactly one of the correct six integers, where the order in which these integers are selected does not matter, from the positive integers not exceeding
 - a) 40. b) 48. c) 56. d) 64.
28. In a superlottery, a player selects 7 numbers out of the first 80 positive integers. What is the probability that a person wins the grand prize by picking 7 numbers that are among the 11 numbers selected at random by a computer.
29. In a superlottery, players win a fortune if they choose the eight numbers selected by a computer from the positive integers not exceeding 100. What is the probability that a player wins this superlottery?
30. What is the probability that a player wins the prize offered for correctly choosing five (but not six) numbers out of six integers chosen at random from the integers between 1 and 40, inclusive?
31. Suppose that 100 people enter a contest and that different winners are selected at random for first, second, and third prizes. What is the probability that Michelle wins one of these prizes if she is one of the contestants?
32. Suppose that 100 people enter a contest and that different winners are selected at random for first, second, and third prizes. What is the probability that Kumar, Janice, and Pedro each win a prize if each has entered the contest?
33. What is the probability that Abby, Barry, and Sylvia win the first, second, and third prizes, respectively, in a drawing if 200 people enter a contest and
 - a) no one can win more than one prize.
 - b) winning more than one prize is allowed.
34. What is the probability that Bo, Colleen, Jeff, and Rohini win the first, second, third, and fourth prizes, respectively, in a drawing if 50 people enter a contest and
 - a) no one can win more than one prize.
 - b) winning more than one prize is allowed.
35. In roulette, a wheel with 38 numbers is spun. Of these, 18 are red, and 18 are black. The other two numbers, which are neither black nor red, are 0 and 00. The probability that when the wheel is spun it lands on any particular number is $1/38$.
 - a) What is the probability that the wheel lands on a red number?
 - b) What is the probability that the wheel lands on a black number twice in a row?
 - c) What is the probability that the wheel lands on 0 or 00?
 - d) What is the probability that in five spins the wheel never lands on either 0 or 00?
 - e) What is the probability that the wheel lands on a number between 1 and 6, inclusive, on one spin, but does not land between them on the next spin?
36. Which is more likely: rolling a total of 8 when two dice are rolled or rolling a total of 8 when three dice are rolled?
37. Which is more likely: rolling a total of 9 when two dice are rolled or rolling a total of 9 when three dice are rolled?
38. Two events E_1 and E_2 are called **independent** if $p(E_1 \cap E_2) = p(E_1)p(E_2)$. For each of the following pairs of events, which are subsets of the set of all possible outcomes when a coin is tossed three times, determine whether or not they are independent.
 - a) E_1 : the first coin comes up tails; E_2 : the second coin comes up heads.

- b) E_1 : the first coin comes up tails; E_2 : two, and not three, heads come up in a row.
- c) E_1 : the second coin comes up tails; E_2 : two, and not three, heads come up in a row.

(We will study independence of events in more depth in Section 6.2.)

- 39. Explain what is wrong with the statement that in the Monty Hall Three-Door Puzzle the probability that the prize is behind the first door you select and the probability that the prize is behind the other of the two doors that Monty does not open are both $1/2$, because there are two doors left.
- 40. Suppose that instead of three doors, there are four doors in the Monty Hall puzzle. What is the probability that you win by not changing once the host, who knows what is

behind each door, opens a losing door and gives you the chance to change doors? What is the probability that you win by changing the door you select to one of the two remaining doors among the three that you did not select?

- 41. This problem was posed by the Chevalier de Méré and was solved by Blaise Pascal and Pierre de Fermat.
 - a) Find the probability of rolling at least one six when a die is rolled four times.
 - b) Find the probability that a double six comes up at least once when a pair of dice is rolled 24 times. Answer the query the Chevalier de Méré made to Pascal asking whether this probability was greater than $1/2$.
 - c) Is it more likely that a six comes up at least once when a die is rolled four times or that a double six comes up at least once when a pair of dice is rolled 24 times?

6.2 Probability Theory

Introduction



In Section 6.1 we introduced the notion of the probability of an event. (Recall that an event is a subset of the possible outcomes of an experiment.) We defined the probability of an event E as Laplace did, that is,

$$p(E) = \frac{|E|}{|S|},$$

the number of outcomes in E divided by the total number of outcomes. This definition assumes that all outcomes are equally likely. However, many experiments have outcomes that are not equally likely. For instance, a coin may be biased so that it comes up heads twice as often as tails. Similarly, the likelihood that the input of a linear search is a particular element in a list, or is not in the list, depends on how the input is generated. How can we model the likelihood of events in such situations? In this section we will show how to define probabilities of outcomes to study probabilities of experiments where outcomes may not be equally likely.

Suppose that a fair coin is flipped four times, and the first time it comes up heads. Given this information, what is the probability that heads comes up three times? To answer this and similar questions, we will introduce the concept of *conditional probability*. Does knowing that the first flip comes up heads change the probability that heads comes up three times? If not, these two events are called *independent*, a concept studied later in this section.

Many questions address a particular numerical value associated with the outcome of an experiment. For instance, when we flip a coin 100 times, what is the probability that exactly 40 heads appear? How many heads should we expect to appear? In this section we will introduce *random variables*, which are functions that associate numerical values to the outcomes of experiments.



HISTORICAL NOTE The Chevalier de Méré was a French nobleman, a famous gambler, and a bon vivant. He was successful at making bets with odds slightly greater than $1/2$ (such as having at least one six come up in four tosses of a die). His correspondence with Pascal asking about the probability of having at least one double six come up when a pair of dice is rolled 24 times led to the development of probability theory. According to one account, Pascal wrote to Fermat about the Chevalier saying something like “He’s a good guy but, alas, he’s not mathematician.”

Assigning Probabilities

Let S be the sample space of an experiment with a finite or countable number of outcomes. We assign a probability $p(s)$ to each outcome s . We require that two conditions be met:

$$(i) \quad 0 \leq p(s) \leq 1 \text{ for each } s \in S$$

and

$$(ii) \quad \sum_{s \in S} p(s) = 1.$$

Condition (i) states that the probability of each outcome is a nonnegative real number no greater than 1. Condition (ii) states that the sum of the probabilities of all possible outcomes should be 1; that is, when we do the experiment, it is a certainty that one of these outcomes occurs. (Note that when the sample space is infinite, $\sum_{s \in S} p(s)$ is a convergent infinite series.) This is a generalization of Laplace's definition in which each of n outcomes is assigned a probability of $1/n$. Indeed, conditions (i) and (ii) are met when Laplace's definition of probabilities of equally likely outcomes is used and S is finite. (See Exercise 4.)

Note that when there are n possible outcomes, x_1, x_2, \dots, x_n , the two conditions to be met are

$$(i) \quad 0 \leq p(x_i) \leq 1 \text{ for } i = 1, 2, \dots, n$$

and

$$(ii) \quad \sum_{i=1}^n p(x_i) = 1.$$

The function p from the set of all outcomes of the sample space S is called a **probability distribution**.

To model an experiment, the probability $p(s)$ assigned to an outcome s should equal the limit of the number of times s occurs divided by the number of times the experiment is performed, as this number grows without bound. (We will assume that all experiments discussed have outcomes that are predictable on the average, so that this limit exists. We also assume that the outcomes of successive trials of an experiment do not depend on past results.)

Remark: We will not discuss probabilities of events when the set of outcomes is not finite or countable, such as when the outcome of an experiment can be any real number. In such cases, integral calculus is usually required for the study of the probabilities of events.

We can model experiments in which outcomes are either equally likely or not equally likely by choosing the appropriate function $p(s)$, as Example 1 illustrates.

EXAMPLE 1 What probabilities should we assign to the outcomes H (heads) and T (tails) when a fair coin is flipped? What probabilities should be assigned to these outcomes when the coin is biased so that heads comes up twice as often as tails?

Solution: For a fair coin, the probability that heads comes up when the coin is flipped equals the probability that tails comes up, so the outcomes are equally likely. Consequently, we assign the probability $1/2$ to each of the two possible outcomes, that is, $p(H) = p(T) = 1/2$.

For the biased coin we have

$$p(H) = 2p(T).$$

Because

$$p(H) + p(T) = 1,$$

it follows that

$$2p(T) + p(T) = 3p(T) = 1.$$

We conclude that $p(T) = 1/3$ and $p(H) = 2/3$. ◀

DEFINITION 1 Suppose that S is a set with n elements. The *uniform distribution* assigns the probability $1/n$ to each element of S .

We now define the probability of an event as the sum of the probabilities of the outcomes in this event.

DEFINITION 2 The *probability* of the event E is the sum of the probabilities of the outcomes in E . That is,

$$p(E) = \sum_{s \in E} p(s).$$

(Note that when E is an infinite set, $\sum_{s \in E} p(s)$ is a convergent infinite series.)

Note that when there are n outcomes in the event E , that is, if $E = \{a_1, a_2, \dots, a_n\}$, then $p(E) = \sum_{i=1}^n p(a_i)$. Note also that the uniform distribution assigns the same probability to an event that Laplace's original definition of probability assigns to this event. The experiment of selecting an element from a sample space with a uniform distribution is called selecting an element of S **at random**.

EXAMPLE 2 Suppose that a die is biased (or loaded) so that 3 appears twice as often as each other number but that the other five outcomes are equally likely. What is the probability that an odd number appears when we roll this die?

Solution: We want to find the probability of the event $E = \{1, 3, 5\}$. By Exercise 2 at the end of this section, we have

$$p(1) = p(2) = p(4) = p(5) = p(6) = 1/7; p(3) = 2/7.$$

It follows that

$$p(E) = p(1) + p(3) + p(5) = 1/7 + 2/7 + 1/7 = 4/7. \quad \blacktriangleleft$$

When events are equally likely and there are a finite number of possible outcomes, the definition of the probability of an event given in this section (Definition 2) agrees with Laplace's definition (Definition 1 of Section 6.1). To see this, suppose that there are n equally likely outcomes; each possible outcome has probability $1/n$, because the sum of their probabilities is 1. Suppose the event E contains m outcomes. According to Definition 2,

$$p(E) = \sum_{i=1}^m \frac{1}{n} = \frac{m}{n}.$$

Because $|E| = m$ and $|S| = n$, it follows that

$$p(E) = \frac{m}{n} = \frac{|E|}{|S|}.$$

This is Laplace's definition of the probability of the event E .

Combinations of Events

The formulae for probabilities of combinations of events in Section 6.1 continue to hold when we use Definition 2 to define the probability of an event. For example, Theorem 1 of Section 6.1 asserts that

$$p(\overline{E}) = 1 - p(E),$$

where \overline{E} is the complementary event of the event E . This equality also holds when Definition 2 is used. To see this, note that because the sum of the probabilities of the n possible outcomes is 1, and each outcome is either in E or in \overline{E} , but not in both, we have

$$\sum_{s \in S} p(s) = 1 = p(E) + p(\overline{E}).$$

Hence, $p(\overline{E}) = 1 - p(E)$.

Under Laplace's definition, by Theorem 2 in Section 6.1, we have

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

whenever E_1 and E_2 are events in a sample space S . This also holds when we define the probability of an event as we do in this section. To see this, note that $p(E_1 \cup E_2)$ is the sum of the probabilities of the outcomes in $E_1 \cup E_2$. When an outcome x is in one, but not both, of E_1 and E_2 , $p(x)$ occurs in exactly one of the sums for $p(E_1)$ and $p(E_2)$. When an outcome x is in both E_1 and E_2 , $p(x)$ occurs in the sum for $p(E_1)$, in the sum for $p(E_2)$, and in the sum for $p(E_1 \cap E_2)$, so it occurs $1 + 1 - 1 = 1$ time on the right-hand side. Consequently, the left-hand side and right-hand side are equal.

The probability of the union of pairwise disjoint events can be found using Theorem 1.

THEOREM 1 If E_1, E_2, \dots is a sequence of pairwise disjoint events in a sample space S , then

$$p\left(\bigcup_i E_i\right) = \sum_i p(E_i).$$

(Note that this theorem applies when the sequence E_1, E_2, \dots consists of a finite number or a countably infinite number of pairwise disjoint events.)

We leave the proof of Theorem 1 to the reader (see Exercises 36 and 37).

Conditional Probability



Suppose that we flip a coin three times, and all eight possibilities are equally likely. Moreover, suppose we know that the event F , that the first flip comes up tails, occurs. Given this information, what is the probability of the event E , that an odd number of tails appears? Because the first flip comes up tails, there are only four possible outcomes: TTT , TTH , THT , and THH , where H and T represent heads and tails, respectively. An odd number of tails appears only for the outcomes TTT and THH . Because the eight outcomes have equal probability, each of the four possible outcomes, given that F occurs, should also have an equal probability of $1/4$. This suggests that we should assign the probability of $2/4 = 1/2$ to E , given that F occurs. This probability is called the **conditional probability** of E given F .

In general, to find the conditional probability of E given F , we use F as the sample space. For an outcome from E to occur, this outcome must also belong to $E \cap F$. With this motivation, we make Definition 3.

DEFINITION 3 Let E and F be events with $p(F) > 0$. The *conditional probability* of E given F , denoted by $p(E | F)$, is defined as

$$p(E | F) = \frac{p(E \cap F)}{p(F)}.$$

EXAMPLE 3



A bit string of length four is generated at random so that each of the 16 bit strings of length four is equally likely. What is the probability that it contains at least two consecutive 0s, given that its first bit is a 0? (We assume that 0 bits and 1 bits are equally likely.)

Solution: Let E be the event that a bit string of length four contains at least two consecutive 0s, and let F be the event that the first bit of a bit string of length four is a 0. The probability that a bit string of length four has at least two consecutive 0s, given that its first bit is a 0, equals

$$p(E | F) = \frac{p(E \cap F)}{p(F)}.$$

Because $E \cap F = \{0000, 0001, 0010, 0011, 0100\}$, we see that $p(E \cap F) = 5/16$. Because there are eight bit strings of length four that start with a 0, we have $p(F) = 8/16 = 1/2$.

Consequently,

$$p(E | F) = \frac{5/16}{1/2} = \frac{5}{8}.$$

EXAMPLE 4 What is the conditional probability that a family with two children has two boys, given they have at least one boy? Assume that each of the possibilities BB , BG , GB , and GG is equally likely, where B represents a boy and G represents a girl. (Note that BG represents a family with an older boy and a younger girl while GB represents a family with an older girl and a younger boy.)

Solution: Let E be the event that a family with two children has two boys, and let F be the event that a family with two children has at least one boy. It follows that $E = \{BB\}$, $F = \{BB, BG, GB\}$, and $E \cap F = \{BB\}$. Because the four possibilities are equally likely, it follows that $p(F) = 3/4$ and $p(E \cap F) = 1/4$. We conclude that

$$p(E | F) = \frac{p(E \cap F)}{p(F)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Independence



Suppose a coin is flipped three times, as described in the introduction to our discussion of conditional probability. Does knowing that the first flip comes up tails (event F) alter the probability that tails comes up an odd number of times (event E)? In other words, is it the case that $p(E | F) = p(E)$? This equality is valid for the events E and F , because $p(E | F) = 1/2$ and $p(E) = 1/2$. Because this equality holds, we say that E and F are **independent events**. When two events are independent, the occurrence of one of the events gives no information about the probability that the other event occurs.

Because $p(E | F) = p(E \cap F)/p(F)$, asking whether $p(E | F) = p(E)$ is the same as asking whether $p(E \cap F) = p(E)p(F)$. This leads to Definition 4.

DEFINITION 4 The events E and F are *independent* if and only if $p(E \cap F) = p(E)p(F)$.

EXAMPLE 5 Suppose E is the event that a randomly generated bit string of length four begins with a 1 and F is the event that this bit string contains an even number of 1s. Are E and F independent, if the 16 bit strings of length four are equally likely?



Solution: There are eight bit strings of length four that begin with a one: 1000, 1001, 1010, 1011, 1100, 1101, 1110, and 1111. There are also eight bit strings of length four that contain an even number of ones: 0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111. Because there are 16 bit strings of length four, it follows that

$$p(E) = p(F) = 8/16 = 1/2.$$

Because $E \cap F = \{1111, 1100, 1010, 1001\}$, we see that

$$p(E \cap F) = 4/16 = 1/4.$$

Because

$$p(E \cap F) = 1/4 = (1/2)(1/2) = p(E)p(F),$$

we conclude that E and F are independent. ◀

Probability has many applications to genetics, as Examples 6 and 7 illustrate.

EXAMPLE 6 Assume, as in Example 4, that each of the four ways a family can have two children is equally likely. Are the events E , that a family with two children has two boys, and F , that a family with two children has at least one boy, independent?

Solution: Because $E = \{BB\}$, we have $p(E) = 1/4$. In Example 4 we showed that $p(F) = 3/4$ and that $p(E \cap F) = 1/4$. But $p(E)p(F) = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$. Therefore $p(E \cap F) \neq p(E)p(F)$, so the events E and F are not independent. ◀

EXAMPLE 7 Are the events E , that a family with three children has children of both sexes, and F , that this family has at most one boy, independent? Assume that the eight ways a family can have three children are equally likely.

Solution: By assumption, each of the eight ways a family can have three children, BBB , BBG , BGB , BGG , GBB , GBG , GGB , and GGG , has a probability of $1/8$. Because $E = \{BBG, BGB, BGG, GBB, GBG, GGB\}$, $F = \{BGG, GBG, GGB, GGG\}$, and $E \cap F = \{BGG, GBG, GGB\}$, it follows that $p(E) = 6/8 = 3/4$, $p(F) = 4/8 = 1/2$, and $p(E \cap F) = 3/8$. Because

$$p(E)p(F) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8},$$

it follows that $p(E \cap F) = p(E)p(F)$, so E and F are independent. (This conclusion may seem surprising. Indeed, if we change the number of children, the conclusion may no longer hold. See Exercise 27 at the end of this section.) ◀

Bernoulli Trials and the Binomial Distribution



Suppose that an experiment can have only two possible outcomes. For instance, when a bit is generated at random, the possible outcomes are 0 and 1. When a coin is flipped, the possible outcomes are heads and tails. Each performance of an experiment with two possible outcomes is called a **Bernoulli trial**, after James Bernoulli, who made important contributions to probability theory. In general, a possible outcome of a Bernoulli trial is called a **success** or a **failure**. If p is the probability of a success and q is the probability of a failure, it follows that $p + q = 1$.

Many problems can be solved by determining the probability of k successes when an experiment consists of n mutually independent Bernoulli trials. (Bernoulli trials are **mutually independent** if the conditional probability of success on any given trial is p , given any information whatsoever about the outcomes of the other trials.) Consider Example 8.

EXAMPLE 8 A coin is biased so that the probability of heads is $2/3$. What is the probability that exactly four heads come up when the coin is flipped seven times, assuming that the flips are independent?

Solution: There are $2^7 = 128$ possible outcomes when a coin is flipped seven times. The number of ways four of the seven flips can be heads is $C(7, 4)$. Because the seven flips are independent,

the probability of each of these outcomes (four heads and three tails) is $(2/3)^4(1/3)^3$. Consequently, the probability that exactly four heads appear is

$$C(7, 4)(2/3)^4(1/3)^3 = \frac{35 \cdot 16}{3^7} = \frac{560}{2187}.$$

◀

Following the same reasoning as was used in Example 8, we can find the probability of k successes in n independent Bernoulli trials.

THEOREM 2 The probability of exactly k successes in n independent Bernoulli trials, with probability of success p and probability of failure $q = 1 - p$, is

$$C(n, k)p^kq^{n-k}.$$

Proof: When n Bernoulli trials are carried out, the outcome is an n -tuple (t_1, t_2, \dots, t_n) , where $t_i = S$ (for success) or $t_i = F$ (for failure) for $i = 1, 2, \dots, n$. Because the n trials are independent, the probability of each outcome of n trials consisting of k successes and $n - k$ failures (in any order) is p^kq^{n-k} . Because there are $C(n, k)$ n -tuples of S 's and F 's that contain k S 's, the probability of k successes is

$$C(n, k)p^kq^{n-k}.$$

◀

We denote by $b(k; n, p)$ the probability of k successes in n independent Bernoulli trials with probability of success p and probability of failure $q = 1 - p$. Considered as a function of k , we call this function the **binomial distribution**. Theorem 2 tells us that $b(k; n, p) = C(n, k)p^kq^{n-k}$.

EXAMPLE 9 Suppose that the probability that a 0 bit is generated is 0.9, that the probability that a 1 bit is generated is 0.1, and that bits are generated independently. What is the probability that exactly eight 0 bits are generated when 10 bits are generated?

Extra
Examples



Solution: By Theorem 2, the probability that exactly eight 0 bits are generated is

$$b(8; 10, 0.9) = C(10, 8)(0.9)^8(0.1)^2 = 0.1937102445.$$

◀

Links



JAMES BERNOULLI (1654–1705) James Bernoulli (also known as Jacob I), was born in Basel, Switzerland. He is one of the eight prominent mathematicians in the Bernoulli family (see Section 10.1 for the Bernoulli family tree of mathematicians). Following his father's wish, James studied theology and entered the ministry. But contrary to the desires of his parents, he also studied mathematics and astronomy. He traveled throughout Europe from 1676 to 1682, learning about the latest discoveries in mathematics and the sciences. Upon returning to Basel in 1682, he founded a school for mathematics and the sciences. He was appointed professor of mathematics at the University of Basel in 1687, remaining in this position for the rest of his life.

James Bernoulli is best known for the work *Ars Conjectandi*, published eight years after his death. In this work, he described the known results in probability theory and in enumeration, often providing alternative proofs of known results. This work also includes the application of probability theory to games of chance and his introduction of the theorem known as the **law of large numbers**. This law states that if $\epsilon > 0$, as n becomes arbitrarily large the probability approaches 1 that the fraction of times an event E occurs during n trials is within ϵ of $p(E)$.

Note that the sum of the probabilities that there are k successes when n independent Bernoulli trials are carried out, for $k = 0, 1, 2, \dots, n$, equals

$$\sum_{k=0}^n C(n, k) p^k q^{n-k} = (p + q)^n = 1,$$

as should be the case. The first equality in this string of equalities is a consequence of the Binomial Theorem (see Section 5.4). The second equality follows because $q = 1 - p$.

Random Variables

Many problems are concerned with a numerical value associated with the outcome of an experiment. For instance, we may be interested in the total number of one bits in a randomly generated string of 10 bits; or in the number of times tails come up when a coin is flipped 20 times. To study problems of this type we introduce the concept of a random variable.

DEFINITION 5 A *random variable* is a function from the sample space of an experiment to the set of real numbers. That is, a random variable assigns a real number to each possible outcome.

Remark: Note that a random variable is a function. It is not a variable, and it is not random!

EXAMPLE 10 Suppose that a coin is flipped three times. Let $X(t)$ be the random variable that equals the number of heads that appear when t is the outcome. Then $X(t)$ takes on the following values:

$$\begin{aligned} X(HHH) &= 3, \\ X(HHT) &= X(HTH) = X(THH) = 2, \\ X(TTH) &= X(THT) = X(HTT) = 1, \\ X(TTT) &= 0. \end{aligned}$$

DEFINITION 6 The *distribution* of a random variable X on a sample space S is the set of pairs $(r, p(X = r))$ for all $r \in X(S)$, where $p(X = r)$ is the probability that X takes the value r . A distribution is usually described by specifying $p(X = r)$ for each $r \in X(S)$.

EXAMPLE 11 Because each of the eight possible outcomes when three coins are flipped has probability $1/8$, the distribution of the random variable $X(t)$ in Example 10 is given by $P(X = 3) = 1/8$, $P(X = 2) = 3/8$, $P(X = 1) = 3/8$, $P(X = 0) = 1/8$.

EXAMPLE 12 Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values of this random variable for the 36 possible outcomes (i, j) , where i and j are the numbers that appear on the first die and the second die, respectively, when these two dice are rolled?

Solution: The random variable X takes on the following values:

$$\begin{aligned}
 X((1, 1)) &= 2, \\
 X((1, 2)) &= X((2, 1)) = 3, \\
 X((1, 3)) &= X((2, 2)) = X((3, 1)) = 4, \\
 X((1, 4)) &= X((2, 3)) = X((3, 2)) = X((4, 1)) = 5, \\
 X((1, 5)) &= X((2, 4)) = X((3, 3)) = X((4, 2)) = X((5, 1)) = 6, \\
 X((1, 6)) &= X((2, 5)) = X((3, 4)) = X((4, 3)) = X((5, 2)) = X((6, 1)) = 7, \\
 X((2, 6)) &= X((3, 5)) = X((4, 4)) = X((5, 3)) = X((6, 2)) = 8, \\
 X((3, 6)) &= X((4, 5)) = X((5, 4)) = X((6, 3)) = 9, \\
 X((4, 6)) &= X((5, 5)) = X((6, 4)) = 10, \\
 X((5, 6)) &= X((6, 5)) = 11, \\
 X((6, 6)) &= 12.
 \end{aligned}$$

We will continue our study of random variables in Section 6.4, where we will show how they can be used in a variety of applications.

The Birthday Problem

A famous puzzle asks for the smallest number of people needed in a room so that it is more likely than not that at least two of them have the same day of the year as their birthday. Most people find the answer, which we determine in Example 13, to be surprisingly small. After we solve this famous problem, we will show how similar reasoning can be adapted to solve a question about hashing functions.

EXAMPLE 13 The Birthday Problem What is the minimum number of people who need to be in a room so that the probability that at least two of them have the same birthday is greater than $1/2$?

Links



Solution: First, we state some assumptions. We assume that the birthdays of the people in the room are independent. Furthermore, we assume that each birthday is equally likely and that there are 366 days in the year. (In reality, more people are born on some days of the year than others, such as days nine months after some holidays including New Year's Eve, and only leap years have 366 days.)

To find the probability that at least two of n people in a room have the same birthday, we first calculate the probability p_n that these people all have different birthdays. Then, the probability that at least two people have the same birthday is $1 - p_n$. To compute p_n , we consider the birthdays of the n people in some fixed order. Imagine them entering the room one at a time; we will compute the probability that each successive person entering the room has a birthday different from those of the people already in the room.

The birthday of the first person certainly does not match the birthday of someone already in the room. The probability that the birthday of the second person is different from that of the first person is $365/366$ because the second person has a different birthday when he or she was born on one of the 365 days of the year other than the day the first person was born. (The assumption that it is equally likely for someone to be born on any of the 366 days of the year enters into this and subsequent steps.)

The probability that the third person has a birthday different from both the birthdays of the first and second people given that these two people have different birthdays is $364/366$. In general, the probability that the j th person, with $2 \leq j \leq 366$, has a birthday different from the

birthdays of the $j - 1$ people already in the room given that these $j - 1$ people have different birthdays is

$$\frac{366 - (j - 1)}{366} = \frac{367 - j}{366}.$$

Because we have assumed that the birthdays of the people in the room are independent, we can conclude that the probability that the n people in the room have different birthdays is

$$p_n = \frac{365}{366} \frac{364}{366} \frac{363}{366} \cdots \frac{367 - n}{366}.$$

It follows that the probability that among n people there are at least two people with the same birthday is

$$1 - p_n = 1 - \frac{365}{366} \frac{364}{366} \frac{363}{366} \cdots \frac{367 - n}{366}.$$

To determine the minimum number of people in the room so that the probability that at least two of them have the same birthday is greater than $1/2$, we use the formula we have found for $1 - p_n$ to compute it for increasing values of n until it becomes greater than $1/2$. (There are more sophisticated approaches using calculus that can eliminate this computation, but we will not use them here.) After considerable computation we find that for $n = 22$, $1 - p_n \approx 0.475$, while for $n = 23$, $1 - p_n \approx 0.506$. Consequently, the minimum number of people needed so that the probability that at least two people have the same birthday is greater than $1/2$ is 23. ◀

The solution to the birthday problem leads to the solution of the question in Example 14 about hashing functions.

EXAMPLE 14 Probability of a Collision in Hashing Functions Recall from Section 3.4 that a hashing function $h(k)$ is a mapping of the keys (of the records that are to be stored in a database) to storage locations. Hashing functions map a large universe of keys (such as the approximately 300 million Social Security numbers in the United States) to a much smaller set of storage locations. A good hashing function yields few **collisions**, which are mappings of two different keys to the same memory location, when relatively few of the records are in play in a given application. What is the probability that no two keys are mapped to the same location by a hashing function, or, in other words, that there are no collisions?

Solution: To calculate this probability, we assume that the probability that a randomly selected key is mapped to a location is $1/m$, where m is the number of available locations, that is, the hashing function distributes keys uniformly. (In practice, hashing functions may not satisfy this assumption. However, for a good hashing function, this assumption should be close to correct.) Furthermore, we assume that the keys of the records selected have an equal probability to be any of the elements of the key universe and that these keys are independently selected.

Suppose that the keys are k_1, k_2, \dots, k_n . When we add the second record, the probability that it is mapped to a location different from the location of the first record, that $h(k_2) \neq h(k_1)$, is $(m - 1)/m$ because there are $m - 1$ free locations after the first record has been placed. The probability that the third record is mapped to a free location after the first and second records have been placed without a collision is $(m - 2)/m$. In general, the probability that the j th record is mapped to a free location after the first $j - 1$ records have been mapped to locations $h(k_1), h(k_2), \dots, h(k_{j-1})$ without collisions is $(m - (j - 1))/m$ because $j - 1$ of the m locations are taken.

Because the keys are independent, the probability that all n keys are mapped to different locations is

$$p_n = \frac{m-1}{m} \cdot \frac{m-2}{m} \cdot \dots \cdot \frac{m-n+1}{m}.$$

It follows that the probability that there is at least one collision, that is, at least two keys are mapped to the same location, is

$$1 - p_n = 1 - \frac{m-1}{m} \cdot \frac{m-2}{m} \cdot \dots \cdot \frac{m-n+1}{m}.$$

Techniques from calculus can be used to find the smallest value of n given a value of m such that the probability of a collision is greater than a particular threshold. It can be shown that the smallest integer n such that the probability of a collision is greater than $1/2$ is approximately $n = 1.177\sqrt{m}$. For example, when $m = 1,000,000$, the smallest integer n such that the probability of a collision is greater than $1/2$ is 1178. ◀

Monte Carlo Algorithms

The algorithms discussed so far in this book are all deterministic. That is, each algorithm always proceeds in the same way whenever given the same input. However, there are many situations where we would like an algorithm to make a random choice at one or more steps. Such a situation arises when a deterministic algorithm would have to go through a huge number, or even an unknown number, of possible cases. Algorithms that make random choices at one or more steps are called **probabilistic algorithms**. We will discuss a particular class of probabilistic algorithms in this section, namely, **Monte Carlo algorithms**, for decision problems. Monte Carlo algorithms always produce answers to problems, but a small probability remains that these answers may be incorrect. However, the probability that the answer is incorrect decreases rapidly when the algorithm carries out sufficient computation. Decision problems have either “true” or “false” as their answer. The designation “Monte Carlo” is a reference to the famous casino in Monaco; the use of randomness and the repetitive processes in these algorithms make them similar to some gambling games. This name was introduced by the inventors of Monte Carlo methods, including Stan Ulam, Enrico Fermi, and John von Neumann.

A Monte Carlo algorithm for a decision problem uses a sequence of tests. The probability that the algorithm answers the decision problem correctly increases as more tests are carried out. At each step of the algorithm, possible responses are “true,” which means that the answer is “true” and no additional iterations are needed, or “unknown,” which means that the answer could be either “true” or “false.” After running all the iterations in such an algorithm, the final answer produced is “true” if at least one iteration yields the answer “true,” and the answer is “false” if every iteration yields the answer “unknown.” If the correct answer is “false,” then the algorithm answers “false,” because every iteration will yield “unknown.” However, if the correct answer is “true,” then the algorithm could answer either “true” or “false,” because it may be possible that each iteration produced the response “unknown” even though the correct response was “true.” We will show that this possibility becomes extremely unlikely as the number of tests increases.

Suppose that p is the probability that the response of a test is “true,” given that the answer is “true.” It follows that $1 - p$ is the probability that the response is “unknown,” given that the answer is “true.” Because the algorithm answers “false” when all n iterations yield the answer “unknown” and the iterations perform independent tests, the probability of error is $(1 - p)^n$. When $p \neq 0$, this probability approaches 0 as the number of tests increases. Consequently, the probability that the algorithm answers “true” when the answer is “true” approaches 1.

EXAMPLE 15 Quality Control (This example is adapted from [AhUI95].) Suppose that a manufacturer orders processor chips in batches of size n , where n is a positive integer. The chip maker has tested only some of these batches to make sure that all the chips in the batch are good (replacing any bad chips found during testing with good ones). In previously untested batches, the probability that a particular chip is bad has been observed to be 0.1 when random testing is done. The PC manufacturer wants to decide whether all the chips in a batch are good. To do this, the PC manufacturer can test each chip in a batch to see whether it is good. However, this requires n tests. Assuming that each test can be carried out in constant time, these tests require $O(n)$ seconds. Can the PC manufacturer test whether a batch of chips has been tested by the chip maker using less time?

Solution: We can use a Monte Carlo algorithm to determine whether a batch of chips has been tested by the chip maker as long as we are willing to accept some probability of error. The algorithm is set up to answer the question: “Has this batch of chips not been tested by the chip maker?” It proceeds by successively selecting chips at random from the batch and testing them one by one. When a bad chip is encountered, the algorithm answers “true” and stops. If a tested chip is good, the algorithm answers “unknown” and goes on to the next chip. After the algorithm has tested a specified number of chips, say k chips, without getting an answer of “true,” the algorithm terminates with the answer “false”; that is, the algorithm concludes that the batch is good, that is, that the chip maker has tested all the chips in the batch.

The only way for this algorithm to answer incorrectly is for it to conclude that an untested batch of chips has been tested by the chip maker. The probability that a chip is good, but that it came from an untested batch, is $1 - 0.1 = 0.9$. Because the events of testing different chips from a batch are independent, the probability that all k steps of the algorithm produce the answer “unknown,” given that the batch of chips is untested, is 0.9^k .

By taking k large enough, we can make this probability as small as we like. For example, by testing 66 chips, the probability that the algorithm decides a batch has been tested by the chip maker is 0.9^{66} , which is less than 0.001. That is, the probability is less than 1 in 1000 that the algorithm has answered incorrectly. Note that this probability is independent of n , the number of chips in a batch. That is, the Monte Carlo algorithm uses a constant number, or $O(1)$, tests and requires $O(1)$ seconds, no matter how many chips are in a batch. As long as the PC manufacturer can live with an error rate of less than 1 in 1000, the Monte Carlo algorithm will save the PC manufacturer a lot of testing. If a smaller error rate is needed, the PC manufacturer can test more chips in each batch; the reader can verify that 132 tests lower the error rate to less than 1 in 1,000,000. ◀

EXAMPLE 16 Probabilistic Primality Testing In Chapter 3 we remarked that a composite integer, that is, an integer that is not prime, passes Miller’s test (see the preamble to Exercise 30 in Section 3.7) for fewer than $n/4$ bases b with $1 < b < n$. This observation is the basis for a Monte Carlo algorithm to determine whether a positive integer is prime. Because large primes play an essential role in public-key cryptography (see Section 3.7), being able to generate large primes quickly has become extremely important.

The goal of the algorithm is to decide the question “Is n composite?” Given an integer n , we select an integer b at random with $1 < b < n$ and determine whether n passes Miller’s test to the base b . If n fails the test, the answer is “true” because n must be composite, and the algorithm ends. Otherwise, we repeat the test k times, where k is an integer. Each time we select a random integer b and determine whether n passes Miller’s test to the base b . If the answer is “unknown” at each step, the algorithm answers “false,” that is, it says that n is not composite, so that it is prime. The only possibility for the algorithm to return an incorrect answer occurs when n is composite, and the answer “unknown” is the output at each of the k iterations. The probability that a composite integer n passes Miller’s test for a randomly selected base b is less than $1/4$. Because the integer b with $1 < b < n$ is selected at random at each iteration and these iterations

are independent, the probability that n is composite but the algorithm responds that n is prime is less than $(1/4)^k$. By taking k to be sufficiently large, we can make this probability extremely small. For example, with 10 iterations, the probability that the algorithm decides that n is prime when it really is composite is less than 1 in 1,000,000. With 30 iterations, this probability drops to less than 1 in 10^{18} , an extremely unlikely event.

To generate large primes, say with 200 digits, we randomly choose an integer n with 200 digits and run this algorithm, with 30 iterations. If the algorithm decides that n is prime, we can use it as one of the two primes used in an encryption key for the RSA cryptosystem. If n is actually composite and is used as part of the key, the procedures used to decrypt messages will not produce the original encrypted message. The key is then discarded and two new possible primes are used. ◀

The Probabilistic Method

We discussed existence proofs in Chapter 1 and illustrated the difference between constructive existence proofs and nonconstructive existence proofs. The probabilistic method, introduced by Paul Erdős and Alfréd Rényi, is a powerful technique that can be used to create nonconstructive existence proofs. To use the probabilistic method to prove results about a set S , such as the existence of an element in S with a specified property, we assign probabilities to the elements of S . We then use methods from probability theory to prove results about the elements of S . In particular, we can show that an element with a specified property exists by showing that the probability an element $x \in S$ has this property is positive. The probabilistic method is based on the equivalent statement in Theorem 3.

THEOREM 3 THE PROBABILISTIC METHOD If the probability that an element of a set S does not have a particular property is less than 1, there exists an element in S with this property.

An existence proof based on the probabilistic method is nonconstructive because it does not find a particular element with the desired property.

We illustrate the power of the probabilistic method by finding a lower bound for the Ramsey number $R(k, k)$. Recall from Section 5.2 that $R(k, k)$ equals the minimum number of people at a party needed to ensure that there are at least k mutual friends or k mutual enemies (assuming that any two people are friends or enemies).

THEOREM 4 If k is an integer with $k \geq 2$, then $R(k, k) \geq 2^{k/2}$.

Proof: We note that the theorem holds for $k = 2$ and $k = 3$ because $R(2, 2) = 2$ and $R(3, 3) = 6$, as was shown in Section 5.2. Now suppose that $k \geq 4$. We will use the probabilistic method to show that if there are fewer than $2^{k/2}$ people at a party, it is possible that no k of them are mutual friends or mutual enemies. This will show that $R(k, k)$ is at least $2^{k/2}$.

To use the probabilistic method, we assume that it is equally likely for two people to be friends or enemies. (Note that this assumption does not have to be realistic.) Suppose there are n people at the party. It follows that there are $\binom{n}{k}$ different sets of k people at this party, which we list as $S_1, S_2, \dots, S_{\binom{n}{k}}$. Let E_i be the event that all k people in S_i are either mutual friends

or mutual enemies. The probability that there are either k mutual friends or k mutual enemies among the n people equals $p(\bigcup_{i=1}^{\binom{n}{k}} E_i)$.

According to our assumption it is equally likely for two people to be friends or enemies. The probability that two people are friends equals the probability that they are enemies; both probabilities equal $1/2$. Furthermore, there are $\binom{k}{2} = k(k-1)/2$ pairs of people in S_i because there are k people in S_i . Hence, the probability that all k people in S_i are mutual friends and the probability that all k people in S_i are mutual enemies both equal $(1/2)^{k(k-1)/2}$. It follows that $p(E_i) = 2(1/2)^{k(k-1)/2}$.

The probability that there are either k mutual friends or k mutual enemies in the group of n people equals $p(\bigcup_{i=1}^{\binom{n}{k}} E_i)$. Using Boole's Inequality (Exercise 15), it follows that

$$p\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq \sum_{i=1}^{\binom{n}{k}} p(E_i) = \binom{n}{k} \cdot 2\left(\frac{1}{2}\right)^{k(k-1)/2}.$$

By Exercise 17 in Section 5.4, we have $\binom{n}{k} \leq n^k/2^{k-1}$. Hence,

$$\binom{n}{k} 2\left(\frac{1}{2}\right)^{k(k-1)/2} \leq \frac{n^k}{2^{k-1}} 2\left(\frac{1}{2}\right)^{k(k-1)/2}.$$

Now if $n < 2^{k/2}$, we have

$$\frac{n^k}{2^{k-1}} 2\left(\frac{1}{2}\right)^{k(k-1)/2} < \frac{2^{k(k/2)}}{2^{k-1}} 2\left(\frac{1}{2}\right)^{k(k-1)/2} = 2^{2-(k/2)} \leq 1$$

where the last step follows because $k \geq 4$.

We can now conclude that $p(\bigcup_{i=1}^{\binom{n}{k}} E_i) < 1$ when $k \geq 4$. Hence, the probability of the complementary event, that there is no set of either k mutual friends or mutual enemies at the party, is greater than 0. It follows that if $n < 2^{k/2}$, there is at least one set such that no subset of k people are mutual friends or mutual enemies. \triangleleft

Exercises

- What probability should be assigned to the outcome of heads when a biased coin is tossed, if heads is three times as likely to come up as tails? What probability should be assigned to the outcome of tails?
- Find the probability of each outcome when a loaded die is rolled, if a 3 is twice as likely to appear as each of the other five numbers on the die.
- Find the probability of each outcome when a biased die is rolled, if rolling a 2 or rolling a 4 is three times as likely as rolling each of the other four numbers on the die and it is equally likely to roll a 2 or a 4.
- Show that conditions (i) and (ii) are met under Laplace's definition of probability, when outcomes are equally likely.
- A pair of dice is loaded. The probability that a 4 appears on the first die is $2/7$, and the probability that a 3 appears on the second die is $2/7$. Other outcomes for each die appear with probability $1/7$. What is the probability of 7 appearing as the sum of the numbers when the two dice are rolled?
- What is the probability of these events when we randomly select a permutation of $\{1, 2, 3\}$?
 - 1 precedes 3.
 - 3 precedes 1.
 - 3 precedes 1 and 3 precedes 2.
- What is the probability of these events when we randomly select a permutation of $\{1, 2, 3, 4\}$?
 - 1 precedes 4.
 - 4 precedes 1.
 - 4 precedes 1 and 4 precedes 2.
 - 4 precedes 1, 4 precedes 2, and 4 precedes 3.
 - 4 precedes 3 and 2 precedes 1.

8. What is the probability of these events when we randomly select a permutation of $\{1, 2, \dots, n\}$ where $n \geq 4$?
- 1 precedes 2.
 - 2 precedes 1.
 - 1 immediately precedes 2.
 - n precedes 1 and $n-1$ precedes 2.
 - n precedes 1 and n precedes 2.
9. What is the probability of these events when we randomly select a permutation of the 26 lowercase letters of the English alphabet?
- The permutation consists of the letters in reverse alphabetic order.
 - z is the first letter of the permutation.
 - z precedes a in the permutation.
 - a immediately precedes z in the permutation.
 - a immediately precedes m , which immediately precedes z in the permutation.
 - m , n , and o are in their original places in the permutation.
10. What is the probability of these events when we randomly select a permutation of the 26 lowercase letters of the English alphabet?
- The first 13 letters of the permutation are in alphabetical order.
 - a is the first letter of the permutation and z is the last letter.
 - a and z are next to each other in the permutation.
 - a and b are not next to each other in the permutation.
 - a and z are separated by at least 23 letters in the permutation.
 - z precedes both a and b in the permutation.
11. Suppose that E and F are events such that $p(E) = 0.7$ and $p(F) = 0.5$. Show that $p(E \cup F) \geq 0.7$ and $p(E \cap F) \geq 0.2$.
12. Suppose that E and F are events such that $p(E) = 0.8$ and $p(F) = 0.6$. Show that $p(E \cup F) \geq 0.8$ and $p(E \cap F) \geq 0.4$.
13. Show that if E and F are events, then $p(E \cap F) \geq p(E) + p(F) - 1$. This is known as **Bonferroni's Inequality**.
14. Use mathematical induction to prove the following generalization of Bonferroni's Inequality:
- $$p(E_1 \cap E_2 \cap \dots \cap E_n) \geq p(E_1) + p(E_2) + \dots + p(E_n) - (n-1),$$
- where E_1, E_2, \dots, E_n are n events.
15. Show that if E_1, E_2, \dots, E_n are events from a finite sample space, then
- $$p(E_1 \cup E_2 \cup \dots \cup E_n) \leq p(E_1) + p(E_2) + \dots + p(E_n).$$
- This is known as **Boole's Inequality**.
16. Show that if E and F are independent events, then \overline{E} and \overline{F} are also independent events.
17. If E and F are independent events, prove or disprove that \overline{E} and F are necessarily independent events.
- In Exercises 18, 20, and 21 assume that the year has 366 days and all birthdays are equally likely. In Exercise 19 assume it is equally likely that a person is born in any given month of the year.
18. a) What is the probability that two people chosen at random were born on the same day of the week?
 b) What is the probability that in a group of n people chosen at random, there are at least two born on the same day of the week?
 c) How many people chosen at random are needed to make the probability greater than $1/2$ that there are at least two people born on the same day of the week?
19. a) What is the probability that two people chosen at random were born during the same month of the year?
 b) What is the probability that in a group of n people chosen at random, there are at least two born in the same month of the year?
 c) How many people chosen at random are needed to make the probability greater than $1/2$ that there are at least two people born in the same month of the year?
20. Find the smallest number of people you need to choose at random so that the probability that at least one of them has a birthday today exceeds $1/2$.
21. Find the smallest number of people you need to choose at random so that the probability that at least two of them were both born on April 1 exceeds $1/2$.
- *22. February 29 occurs only in leap years. Years divisible by 4, but not by 100, are always leap years. Years divisible by 100, but not by 400, are not leap years, but years divisible by 400 are leap years.
- What probability distribution for birthdays should be used to reflect how often February 29 occurs?
 - Using the probability distribution from part (a), what is the probability that in a group of n people at least two have the same birthday?
23. What is the conditional probability that exactly four heads appear when a fair coin is flipped five times, given that the first flip came up heads?
24. What is the conditional probability that exactly four heads appear when a fair coin is flipped five times, given that the first flip came up tails?
25. What is the conditional probability that a randomly generated bit string of length four contains at least two consecutive 0s, given that the first bit is a 1? (Assume the probabilities of a 0 and a 1 are the same.)
26. Let E be the event that a randomly generated bit string of length three contains an odd number of 1s, and let F be the event that the string starts with 1. Are E and F independent?
27. Let E and F be the events that a family of n children has children of both sexes and has at most one boy, respectively. Are E and F independent if
- $n = 2$?
 - $n = 4$?
 - $n = 5$?

28. Assume that the probability a child is a boy is 0.51 and that the sexes of children born into a family are independent. What is the probability that a family of five children has
- exactly three boys?
 - at least one boy?
 - at least one girl?
 - all children of the same sex?
29. A group of six people play the game of “odd person out” to determine who will buy refreshments. Each person flips a fair coin. If there is a person whose outcome is not the same as that of any other member of the group, this person has to buy the refreshments. What is the probability that there is an odd person out after the coins are flipped once?
30. Find the probability that a randomly generated bit string of length 10 does not contain a 0 if bits are independent and if
- a 0 bit and a 1 bit are equally likely.
 - the probability that a bit is a 1 is 0.6.
 - the probability that the i th bit is a 1 is $1/2^i$ for $i = 1, 2, 3, \dots, 10$.
31. Find the probability that a family with five children does not have a boy, if the sexes of children are independent and if
- a boy and a girl are equally likely.
 - the probability of a boy is 0.51.
 - the probability that the i th child is a boy is $0.51 - (i/100)$.
32. Find the probability that a randomly generated bit string of length 10 begins with a 1 or ends with a 00 for the same conditions as in parts (a), (b), and (c) of Exercise 30, if bits are generated independently.
33. Find the probability that the first child of a family with five children is a boy or that the last two children of the family are girls, for the same conditions as in parts (a), (b), and (c) of Exercise 31.
34. Find each of the following probabilities when n independent Bernoulli trials are carried out with probability of success p .
- the probability of no successes
 - the probability of at least one success
 - the probability of at most one success
 - the probability of at least two successes
35. Find each of the following probabilities when n independent Bernoulli trials are carried out with probability of success p .
- the probability of no failures
 - the probability of at least one failure
 - the probability of at most one failure
 - the probability of at least two failures
36. Use mathematical induction to prove that if E_1, E_2, \dots, E_n is a sequence of n pairwise disjoint events in a sample space S , where n is a positive integer, then $p(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n p(E_i)$.
- *37. (Requires calculus) Show that if E_1, E_2, \dots is an infinite sequence of pairwise disjoint events in a sample space S , then $p(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} p(E_i)$. [Hint: Use Exercise 36 and take limits.]
38. A pair of dice is rolled in a remote location and when you ask an honest observer whether at least one die came up six, this honest observer answers in the affirmative.
- What is the probability that the sum of the numbers that came up on the two dice is seven, given the information provided by the honest observer?
 - Suppose that the honest observer tells us that at least one die came up five. What is the probability the sum of the numbers that came up on the dice is seven, given this information?
- **39. This exercise employs the probabilistic method to prove a result about round-robin tournaments. In a **round-robin tournament** with m players, every two players play one game in which one player wins and the other loses.
- We want to find conditions on positive integers m and k with $k < m$ such that it is possible for the outcomes of the tournament to have the property that for every set of k players, there is a player who beats every member in this set. So that we can use probabilistic reasoning to draw conclusions about round-robin tournaments, we assume that when two players compete it is equally likely that either player wins the game and we assume that the outcomes of different games are independent. Let E be the event that for every set S with k players, where k is a positive integer less than m , there is a player who has beaten all k players in S .
- Show that $p(\bar{E}) \leq \sum_{j=1}^{\binom{m}{k}} p(F_j)$, where F_j is the event that there is no player who beats all k players from the j th set in a list of the $\binom{m}{k}$ sets of k players.
 - Show that the probability of F_j is $(1 - 2^{-k})^{m-k}$.
 - Conclude from parts (a) and (b) that $p(\bar{E}) \leq \binom{m}{k}(1 - 2^{-k})^{m-k}$ and, therefore, that there must be a tournament with the described property if $\binom{m}{k}(1 - 2^{-k})^{m-k} < 1$.
 - Use part (c) to find values of m such that there is a tournament with m players such that for every set S of two players, there is a player who has beaten both players in S . Repeat for sets of three players.
- *40. Devise a Monte Carlo algorithm that determines whether a permutation of the integers 1 through n has already been sorted (that is, it is in increasing order), or instead, is a random permutation. A step of the algorithm should answer “true” if it determines the list is not sorted and “unknown” otherwise. After k steps, the algorithm decides that the integers are sorted if the answer is “unknown” in each step. Show that as the number of steps increases, the probability that the algorithm produces an incorrect answer is extremely small. [Hint: For each step, test whether certain elements are in the correct order. Make sure these tests are independent.]
41. Use pseudocode to write out the probabilistic primality test described in Example 16.

6.3 Bayes' Theorem

Introduction

There are many times when we want to assess the probability that a particular event occurs on the basis of partial evidence. For example, suppose we know the percentage of people who have a particular disease for which there is a very accurate diagnostic test. People who test positive for this disease would like to know the likelihood that they actually have the disease. In this section we introduce a result that can be used to determine this probability, namely, the probability that a person has the disease given that they test positive for it. To use this result, we will need to know the percentage of people who do not have the disease but test positive for it and the percentage of people who have the disease but test negative for it.

Similarly, suppose we know the percentage of incoming e-mail messages that are spam. We will see that we can determine the likelihood that an incoming e-mail message is spam using the occurrence of words in the message. To determine this likelihood, we need to know the percentage of incoming messages that are spam, the percentage of spam messages in which each of these words occurs, and the percentage of messages that are not spam in which each of these words occurs.

The result that we can use to answer questions such as these is called Bayes' Theorem and dates back to the eighteenth century. In the past two decades, Bayes' Theorem has been extensively applied to estimate probabilities based on partial evidence in areas as diverse as medicine, law, machine learning, engineering, and software development.

Bayes' Theorem

We illustrate the idea behind Bayes' Theorem with an example that shows that when extra information is available, we can derive a more realistic estimate that a particular event occurs. That is, suppose we know $p(F)$, the probability that an event F occurs, but we have knowledge that an event E occurs. Then the conditional probability that F occurs given that E occurs, $p(F | E)$, is a more realistic estimate than $p(F)$ that F occurs. In Example 1 we will see that we can find $p(F | E)$ when we know $p(F)$, $p(E | F)$, and $p(E | \bar{F})$.

EXAMPLE 1 We have two boxes. The first contains two green balls and seven red balls; the second contains four green balls and three red balls. Bob selects a ball by first choosing one of the two boxes at random. He then selects one of the balls in this box at random. If Bob has selected a red ball, what is the probability that he selected a ball from the first box?



Solution: Let E be the event that Bob has chosen a red ball; \bar{E} is the event that Bob has chosen a green ball. Let F be the event that Bob has chosen a ball from the first box; \bar{F} is the event that Bob has chosen a ball from the second box. We want to find $p(F | E)$, the probability that the ball Bob selected came from the first box, given that it is red. By the definition of conditional probability, we have $p(F | E) = p(F \cap E)/p(E)$. Can we use the information provided to determine both $p(F \cap E)$ and $p(E)$ so that we can find $p(F | E)$?

First, note that because the first box contains seven red balls out of a total of nine balls, we know that $p(E | F) = 7/9$. Similarly, because the second box contains three red balls out of a total of seven balls, we know that $p(E | \bar{F}) = 3/7$. We assumed that Bob selects a box at random, so $p(F) = p(\bar{F}) = 1/2$. Because $p(E | F) = p(E \cap F)/p(F)$, it follows that $p(E \cap F) = p(E | F)p(F) = \frac{7}{9} \cdot \frac{1}{2} = \frac{7}{18}$ [as we remarked earlier, this is one of the quantities

we need to find to determine $p(F | E)$. Similarly, because $p(E | \bar{F}) = p(E \cap \bar{F})/p(\bar{F})$, it follows that $p(E \cap \bar{F}) = p(E | \bar{F})p(\bar{F}) = \frac{3}{7} \cdot \frac{1}{2} = \frac{3}{14}$.

We can now find $p(E)$. Note that $E = (E \cap F) \cup (E \cap \bar{F})$, where $E \cap F$ and $E \cap \bar{F}$ are disjoint sets. (If x belongs to both $E \cap F$ and $E \cap \bar{F}$, then x belongs to both F and \bar{F} , which is impossible.) It follows that

$$p(E) = p(E \cap F) + p(E \cap \bar{F}) = \frac{7}{18} + \frac{3}{14} = \frac{49}{126} + \frac{27}{126} = \frac{76}{126} = \frac{38}{63}.$$

We have now found both $p(F \cap E) = 7/18$ and $p(E) = 38/63$. We conclude that

$$p(F | E) = \frac{p(F \cap E)}{p(E)} = \frac{7/18}{38/63} = \frac{49}{76} \approx 0.645.$$

Before we had any extra information, we assumed that the probability that Bob selected the first box was $1/2$. However, with the extra information that the ball selected at random is red, this probability has increased to approximately 0.645. That is, the probability that Bob selected a ball from the first box increased from $1/2$, when no extra information was available, to 0.645 once we knew that the ball selected was red. ◀

Using the same type of reasoning as in Example 1, we can find the conditional probability that an event F occurs, given that an event E has occurred, when we know $p(E | F)$, $p(E | \bar{F})$, and $p(F)$. The result we can obtain is called **Bayes' Theorem**; it is named after Thomas Bayes, an eighteenth-century British mathematician and minister who introduced this result.

THEOREM 1 BAYES' THEOREM Suppose that E and F are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then

$$p(F | E) = \frac{p(E | F)p(F)}{p(E | F)p(F) + p(E | \bar{F})p(\bar{F})}.$$

Proof: The definition of conditional probability tells us that $p(F | E) = p(E \cap F)/p(E)$ and $p(E | F) = p(E \cap F)/p(F)$. Therefore, $p(E \cap F) = p(F | E)p(E)$ and $p(E \cap F) = p(E | F)p(F)$. Equating these two expressions for $p(E \cap F)$ shows that

$$p(F | E)p(E) = p(E | F)p(F).$$

Dividing both sides by $p(E)$, we find that

$$p(F | E) = \frac{p(E | F)p(F)}{p(E)}.$$

To complete the proof, we show that $p(E) = p(E | F)p(F) + p(E | \bar{F})p(\bar{F})$. First, note that $E = E \cap S = E \cap (F \cup \bar{F}) = (E \cap F) \cup (E \cap \bar{F})$. Furthermore, $E \cap F$ and $E \cap \bar{F}$ are disjoint, because if $x \in E \cap F$ and $x \in E \cap \bar{F}$, then $x \in F \cap \bar{F} = \emptyset$. Consequently, $p(E) =$

$p(E \cap F) + p(E \cap \bar{F})$. We have already shown that $p(E \cap F) = p(E | F)p(F)$. Moreover, we have $p(E | \bar{F}) = p(E \cap \bar{F})/p(\bar{F})$, which shows that $p(E \cap \bar{F}) = p(E | \bar{F})p(\bar{F})$. It follows that

$$p(E) = p(E \cap F) + p(E \cap \bar{F}) = p(E | F)p(F) + p(E | \bar{F})p(\bar{F}).$$

To complete the proof we insert this expression for $p(E)$ into the equation $p(F | E) = p(E | F)p(F)/p(E)$. We have proved that



$$p(F | E) = \frac{p(E | F)p(F)}{p(E | F)p(F) + p(E | \bar{F})p(\bar{F})}.$$



Bayes' Theorem can be used to assess the probability that someone testing positive for a disease actually has this disease. The results obtained from Bayes' Theorem are often somewhat surprising, as Example 2 shows.

EXAMPLE 2 Suppose that one person in 100,000 has a particular rare disease for which there is a fairly accurate diagnostic test. This test is correct 99% of the time when given to someone with the disease; it is correct 99.5% of the time when given to someone who does not have the disease. Given this information can we find

- (a) the probability that someone who tests positive for the disease has the disease?
- (b) the probability that someone who tests negative for the disease does not have the disease?

Should someone who tests positive be very concerned that he or she has the disease?

Solution: (a) Let F be the event that a person has the disease, and let E be the event that this person tests positive for the disease. We want to compute $p(F | E)$. To use Bayes' Theorem to compute $p(F | E)$ we need to find $p(E | F)$, $p(E | \bar{F})$, $p(F)$, and $p(\bar{F})$.

We know that one person in 100,000 has this disease, so that $p(F) = 1/100,000 = 0.00001$ and $p(\bar{F}) = 1 - 0.00001 = 0.99999$. Because someone who has the disease tests positive 99% of the time, we know that $p(E | F) = 0.99$; this is the probability of a true positive, that someone with the disease tests positive. We also know that $p(\bar{E} | F) = 0.01$; this is the probability of a false negative, that someone who has the disease tests negative. Furthermore, because someone who does not have the disease tests negative 99.5% of the time, we know that $p(\bar{E} | \bar{F}) = 0.995$. This is the probability of a true negative, that someone without the disease tests negative. Finally, we know that $p(E | \bar{F}) = 0.005$; this is the probability of a false positive, that someone without the disease tests positive.

The probability that someone who tests positive for the disease actually has the disease is $p(F | E)$. By Bayes' Theorem, we know that

$$\begin{aligned} p(F | E) &= \frac{p(E | F)p(F)}{p(E | F)p(F) + p(E | \bar{F})p(\bar{F})} \\ &= \frac{(0.99)(0.00001)}{(0.99)(0.00001) + (0.005)(0.99999)} \approx 0.002. \end{aligned}$$

This means that only 0.2% of people who test positive for the disease actually have the disease. Because the disease is extremely rare, the number of false positives on the diagnostic test is far greater than the number of true positives, making the percentage of people who test positive

who actually have the disease extremely small. People who test positive for the diseases should not be overly concerned that they actually have the disease.

(b) The probability that someone who tests negative for the disease does not have the disease is $p(\bar{F} | \bar{E})$. By Bayes' Theorem, we know that

$$\begin{aligned} p(\bar{F} | \bar{E}) &= \frac{p(\bar{E} | \bar{F})p(\bar{F})}{p(\bar{E} | \bar{F})p(\bar{F}) + p(\bar{E} | F)p(F)} \\ &= \frac{(0.995)(0.99999)}{(0.995)(0.99999) + (0.01)(0.00001)} \approx 0.9999999. \end{aligned}$$

Consequently, 99.99999% of the people who test negative really do not have the disease. ◀

Note that in the statement of Bayes' Theorem, the events F and \bar{F} are mutually exclusive and cover the entire sample space S (that is, $F \cup \bar{F} = S$). We can extend Bayes' Theorem to any collection of mutually exclusive events that cover the entire sample space S , in the following way.

THEOREM 2 GENERALIZED BAYES' THEOREM Suppose that E is an event from a sample space S and that F_1, F_2, \dots, F_n are mutually exclusive events such that $\bigcup_{i=1}^n F_i = S$. Assume that $p(E) \neq 0$ and $p(F_i) \neq 0$ for $i = 1, 2, \dots, n$. Then

$$p(F_j | E) = \frac{p(E | F_j)p(F_j)}{\sum_{i=1}^n p(E | F_i)p(F_i)}.$$

We leave the proof of this generalized version of Bayes' Theorem as Exercise 17.



THOMAS BAYES (1702–1761) was the son of a minister in a religious sect known as the Nonconformists. This sect was considered heretical in eighteenth-century Great Britain. Because of the secrecy of the Nonconformists, little is known of Thomas Bayes' life. When Thomas was young, his family moved to London. Thomas was likely educated privately; Nonconformist children generally did not attend school. In 1719 Bayes entered the University of Edinburgh, where he studied logic and theology. He was ordained as a Nonconformist minister like his father and began his work as a minister assisting his father. In 1733 he became minister of the Presbyterian Chapel in Tunbridge Wells, southeast of London, where he remained minister until 1752.

Bayes is best known for his essay on probability published in 1764, three years after his death. This essay was sent to the Royal Society by a friend who found it in the papers left behind when Bayes died. In the introduction to this essay, Bayes stated that his goal was to find a method that could measure the probability that an event happens, assuming that we know nothing about it, but that, under the same circumstances, it has happened a certain proportion of times. Bayes' conclusions were accepted by the great French mathematician Laplace but were later challenged by Boole, who questioned them in his book *Laws of Thought*. Since then Bayes' techniques have been subject to controversy.

Bayes also wrote an article that was published posthumously: "An Introduction to the Doctrine of Fluxions, and a Defense of the Mathematicians Against the Objections of the Author of The Analyst," which supported the logical foundations of calculus. Bayes was elected a Fellow of the Royal Society in 1742, with the support of important members of the Society, even though at that time he had no published mathematical works. Bayes' sole known publication during his lifetime was allegedly a mystical book entitled *Divine Benevolence*, discussing the original causation and ultimate purpose of the universe. Although the book is commonly attributed to Bayes, no author's name appeared on the title page, and the entire work is thought to be of dubious provenance. Evidence for Bayes' mathematical talents comes from a notebook that was almost certainly written by Bayes, which contains much mathematical work, including discussions of probability, trigonometry, geometry, solutions of equations, series, and differential calculus. There are also sections on natural philosophy, in which Bayes looks at topics that include electricity, optics, and celestial mechanics. Bayes is also the author of a mathematical publication on asymptotic series, which appeared after his death.

Bayesian Spam Filters



Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**. Because spam threatens to overwhelm electronic mail systems, a tremendous amount of work has been devoted to filtering it out. Some of the first tools developed for eliminating spam were based on Bayes' Theorem, such as **Bayesian spam filters**. A Bayesian spam filter uses information about previously seen e-mail messages to guess whether an incoming e-mail message is spam. Bayesian spam filters look for occurrences of particular words in messages. For a particular word w , the probability that w appears in a spam e-mail message is estimated by determining the number of times w appears in a message from a large set of messages known to be spam and the number of times it appears in a large set of messages known not to be spam. When we examine e-mail messages to determine whether they might be spam, we look at words that might be indicators of spam, such as “offer,” “special,” or “opportunity,” as well as words that might indicate that a message is not spam, such as “mom,” “lunch,” or “Jan” (where Jan is one of your friends). Unfortunately, spam filters sometimes fail to identify a spam message as spam; this is called a false negative. And they sometimes identify a message that is not spam as spam; this is called a false positive. When testing for spam, it is important to minimize false positives, because filtering out wanted e-mail is much worse than letting some spam through.

We will develop some basic Bayesian spam filters. First, suppose we have a set B of messages known to be spam and a set G of messages known not to be spam. (For example, users could classify messages as spam when they examine them in their inboxes.) We next identify the words that occur in B and in G . We count the number of messages in the set containing each word to find $n_B(w)$ and $n_G(w)$, the number of messages containing the word w in the sets B and G , respectively. Then, the empirical probability that a spam message contains the word w is $p(w) = n_B(w)/|B|$, and the empirical probability that a message that is not spam contains the word w is $q(w) = n_G(w)/|G|$. We note that $p(w)$ and $q(w)$ estimate the probabilities that an incoming spam message, and an incoming message that is not spam, contain the word w , respectively.

Now suppose we receive a new e-mail message containing the word w . Let S be the event that the message is spam. Let E be the event that the message contains the word w . The events S , that the message is spam, and \bar{S} , that the message is not spam, partition the set of all messages. Hence, by Bayes' Theorem, the probability that the message is spam, given that it contains the word w , is

$$p(S | E) = \frac{p(E | S)p(S)}{p(E | S)p(S) + p(E | \bar{S})p(\bar{S})}.$$

To apply this formula, we first estimate $p(S)$, the probability that an incoming message is spam, as well as $p(\bar{S})$, the probability that the incoming message is not spam. Without prior knowledge about the likelihood that an incoming message is spam, for simplicity we assume that the message is equally likely to be spam as it is not to be spam. That is, we assume that $p(S) = p(\bar{S}) = 1/2$. Using this assumption, we find that the probability that a message is spam, given that it contains the word w , is

$$p(S | E) = \frac{p(E | S)}{p(E | S) + p(E | \bar{S})}.$$

(Note that if we have some empirical data about the ratio of spam messages to messages that are not spam, we can change this assumption to produce a better estimate for $p(S)$ and for $p(\bar{S})$; see Exercise 22.)

Next, we estimate $p(E | S)$, the conditional probability that the message contains the word w given that the message is spam, by $p(w)$. Similarly, we estimate $p(E | \bar{S})$, the conditional

probability that the message contains the word w , given that the message is not spam, by $q(w)$. Inserting these estimates for $p(E | S)$ and $p(E | \bar{S})$ tells us that $p(S | E)$ can be estimated by

$$r(w) = \frac{p(w)}{p(w) + q(w)};$$

that is, $r(w)$ estimates the probability that the message is spam, given that it contains the word w . If $r(w)$ is greater than a threshold that we set, such as 0.9, then we classify the message as spam.

EXAMPLE 3 Suppose that we have found that the word “Rolex” occurs in 250 of 2000 messages known to be spam and in 5 of 1000 messages known not to be spam. Estimate the probability that an incoming message containing the word “Rolex” is spam, assuming that it is equally likely that an incoming message is spam or not spam. If our threshold for rejecting a message as spam is 0.9, will we reject this message?

Solution: We use the counts that the word “Rolex” appears in spam messages and messages that are not spam to find that $r(\text{Rolex}) = 250/2000 = 0.125$ and $q(\text{Rolex}) = 5/1000 = 0.005$. Because we are assuming that it is equally likely that an incoming message is or is not spam, we can estimate the probability that the message is spam by

$$r(\text{Rolex}) = \frac{p(\text{Rolex})}{p(\text{Rolex}) + q(\text{Rolex})} = \frac{0.125}{0.125 + 0.005} = \frac{0.125}{0.130} \approx 0.962.$$

Because $r(\text{Rolex})$ is greater than the threshold 0.9, we reject this message as spam. ◀

Detecting spam based on the presence of a single word can lead to excessive false positives and false negatives. Consequently, spam filters look at the presence of multiple words. For example, suppose that the message contains the words w_1 and w_2 . Let E_1 and E_2 denote the events that the message contains the words w_1 and w_2 , respectively. To make our computations simpler, we assume that E_1 and E_2 are independent events and that $E_1 | S$ and $E_2 | S$ are independent events and that we have no prior knowledge regarding whether or not the message is spam. (The assumptions that E_1 and E_2 are independent and that $E_1 | S$ and $E_2 | S$ are independent may introduce some error into our computations; we assume that this error is small.) Using Bayes’ Theorem and our assumptions, we can show (see Exercise 23) that $p(S | E_1 \cap E_2)$, the probability that the message is spam given that it contains both w_1 and w_2 , is

$$p(S | E_1 \cap E_2) = \frac{p(E_1 | S)p(E_2 | S)}{p(E_1 | S)p(E_2 | S) + p(E_1 | \bar{S})p(E_2 | \bar{S})}.$$

We estimate the probability $p(S | E_1 \cap E_2)$ by

$$r(w_1, w_2) = \frac{p(w_1)p(w_2)}{p(w_1)p(w_2) + q(w_1)q(w_2)}.$$


That is, $r(w_1, w_2)$ estimates the probability that the message is spam, given that it contains the words w_1 and w_2 . When $r(w_1, w_2)$ is greater than a preset threshold, such as 0.9, we determine that the message is likely spam.

EXAMPLE 4 Suppose that we train a Bayesian spam filter on a set of 2000 spam messages and 1000 messages that are not spam. The word “stock” appears in 400 spam messages and 60 messages that are not spam, and the word “undervalued” appears in 200 spam messages and 25 messages that are

not spam. Estimate the probability that an incoming message containing both the words “stock” and “undervalued” is spam, assuming that we have no prior knowledge about whether it is spam. Will we reject the message as spam when we set the threshold at 0.9?

Solution: Using the counts of each of these two words in messages known to be spam or known not to be spam, we obtain the following estimates: $p(\text{stock}) = 400/2000 = 0.2$, $q(\text{stock}) = 60/1000 = 0.06$, $p(\text{undervalued}) = 200/2000 = 0.1$, and $q(\text{undervalued}) = 25/1000 = 0.025$. Using these probabilities, we can estimate the probability that the message is spam by

$$\begin{aligned} r(\text{stock, undervalued}) &= \frac{p(\text{stock})p(\text{undervalued})}{p(\text{stock})p(\text{undervalued}) + q(\text{stock})q(\text{undervalued})} \\ &= \frac{(0.2)(0.1)}{(0.2)(0.1) + (0.06)(0.025)} \approx 0.930. \end{aligned}$$

Because we have set the threshold for rejecting messages at 0.9, the message will be rejected by the filter. 

The more words we use to estimate the probability that an incoming mail message is spam, the better is our chance that we correctly determine whether it is spam. In general, if E_i is the event that the message contains word w_i , assuming that the number of incoming spam messages is approximately the same as the number of incoming messages that are not spam, and that the events $E_i \mid S$ are independent, then by Bayes' Theorem the probability that a message containing all the words w_i , $i = 1, 2, \dots, k$, is

$$p(S \mid \bigcap_{i=1}^k E_i) = \frac{\prod_{i=1}^k p(E_i \mid S)}{\prod_{i=1}^k p(E_i \mid S) + \prod_{i=1}^k p(E_i \mid \bar{S})}.$$

We can estimate this probability by

$$r(w_1, w_2, \dots, w_k) = \frac{\prod_{i=1}^k p(w_i)}{\prod_{i=1}^k p(w_i) + \prod_{i=1}^k q(w_i)}.$$

For the most effective spam filter, we choose words for which the probability that each of these words appears in spam is either very high or very low. When we compute this value for a particular message, we reject the message as spam if $r(w_1, w_2, \dots, w_k)$ exceeds a preset threshold, such as 0.9.

Another way to improve the performance of a Bayesian spam filter is to look at the probabilities that particular pairs of words appear in spam and in messages that are not spam. We then treat appearances of these pairs of words as appearance of a single block, rather than as the appearance of two separate words. For example, the pair of words “enhance performance” most likely indicates spam, while “operatic performance” indicates a message that is not spam. Similarly, we can assess the likelihood that a message is spam by examining the structure of a message to determine where words appear in it. Also, spam filters look at appearances of certain types of strings of characters rather than just words. For example, a message with the valid e-mail address of one of your friends is less likely to be spam (if not sent by a worm) than one containing an e-mail address that came from a country known to originate a lot of spam. There is an ongoing war between people who create spam and those trying to filter their messages out. This leads to the introduction of many new techniques to defeat spam filters, including inserting into spam messages long strings of words that appear in messages that are not spam, as well as including words inside pictures. The techniques we have discussed here are only the first steps in fighting this war on spam.

Exercises

1. Suppose that E and F are events in a sample space and $p(E) = 1/3$, $p(F) = 1/2$, and $p(E | F) = 2/5$. Find $p(F | E)$.
2. Suppose that E and F are events in a sample space and $p(E) = 2/3$, $p(F) = 3/4$, and $p(F | E) = 5/8$. Find $p(E | F)$.
3. Suppose that Frida selects a ball by first picking one of two boxes at random and then selecting a ball from this box at random. The first box contains two white balls and three blue balls, and the second box contains four white balls and one blue ball. What is the probability that Frida picked a ball from the first box if she has selected a blue ball?
4. Suppose that Ann selects a ball by first picking one of two boxes at random and then selecting a ball from this box. The first box contains three orange balls and four black balls, and the second box contains five orange balls and six black balls. What is the probability that Ann picked a ball from the second box if she has selected an orange ball?
5. Suppose that 8% of all bicycle racers use steroids, that a bicyclist who uses steroids tests positive for steroids 96% of the time, and that a bicyclist who does not use steroids tests positive for steroids 9% of the time. What is the probability that a randomly selected bicyclist who tests positive for steroids actually uses steroids?
6. When a test for steroids is given to soccer players, 98% of the players taking steroids test positive and 12% of the players not taking steroids test positive. Suppose that 5% of soccer players take steroids. What is the probability that a soccer player who tests positive takes steroids?
7. Suppose that a test for opium use has a 2% false positive rate and a 5% false negative rate. That is, 2% of people who do not use opium test positive for opium, and 5% of opium users test negative for opium. Furthermore, suppose that 1% of people actually use opium.
 - a) Find the probability that someone who tests negative for opium use does not use opium.
 - b) Find the probability that someone who tests positive for opium use actually uses opium.
8. Suppose that one person in 10,000 people has a rare genetic disease. There is an excellent test for the disease; 99.9% of people with the disease test positive and only 0.02% who do not have the disease test positive.
 - a) What is the probability that someone who tests positive has the genetic disease?
 - b) What is the probability that someone who tests negative does not have the disease?
9. Suppose that 8% of the patients tested in a clinic are infected with HIV. Furthermore, suppose that when a blood test for HIV is given, 98% of the patients infected with HIV test positive and that 3% of the patients not infected with HIV test positive. What is the probability that
 - a) a patient testing positive for HIV with this test is infected with it?
 - b) a patient testing positive for HIV with this test is not infected with it?
 - c) a patient testing negative for HIV with this test is infected with it?
 - d) a patient testing negative for HIV with this test is not infected with it?
10. Suppose that 4% of the patients tested in a clinic are infected with avian influenza. Furthermore, suppose that when a blood test for avian influenza is given, 97% of the patients infected with avian influenza test positive and that 2% of the patients not infected with avian influenza test positive. What is the probability that
 - a) a patient testing positive for avian influenza with this test is infected with it?
 - b) a patient testing positive for avian influenza with this test is not infected with it?
 - c) a patient testing negative for avian influenza with this test is infected with it?
 - d) a patient testing negative for avian influenza with this test is not infected with it?
11. An electronics company is planning to introduce a new camera phone. The company commissions a marketing report for each new product that predicts either the success or the failure of the product. Of new products introduced by the company, 60% have been successes. Furthermore, 70% of their successful products were predicted to be successes, while 40% of failed products were predicted to be successes. Find the probability this new camera phone will be successful if its success has been predicted.
- *12. A space probe near Neptune communicates with Earth using bit strings. Suppose that in its transmissions it sends a 1 one-third of the time and a 0 two-thirds of the time. When a 0 is sent, the probability that it is received correctly is 0.9, and the probability that it is received incorrectly (as a 1) is 0.1. When a 1 is sent, the probability that it is received correctly is 0.8, and the probability that it is received incorrectly (as a 0) is 0.2.
 - a) Find the probability that a 0 is received.
 - b) Use Bayes' Theorem to find the probability that a 0 was transmitted, given that a 0 was received.
13. Suppose that E , F_1 , F_2 , and F_3 are events from a sample space S and that F_1 , F_2 , and F_3 are mutually disjoint and their union is S . Find $p(F_1 | E)$ if $p(E | F_1) = 1/8$, $p(E | F_2) = 1/4$, $p(E | F_3) = 1/6$, $p(F_1) = 1/4$, $p(F_2) = 1/4$, and $p(F_3) = 1/2$.
14. Suppose that E , F_1 , F_2 , and F_3 are events from a sample space S and that F_1 , F_2 , and F_3 are mutually disjoint and their union is S . Find $p(F_2 | E)$ if $p(E | F_1) = 2/7$,

$p(E | F_2) = 3/8$, $p(E | F_3) = 1/2$, $p(F_1) = 1/6$, $p(F_2) = 1/2$, and $p(F_3) = 1/3$.

15. In this exercise we will use Bayes' Theorem to solve the Monty Hall puzzle (Example 10 in Section 6.1). Recall that in this puzzle you are asked to select one of three doors to open. There is a large prize behind one of the three doors and the other two doors are losers. After you select a door, Monty Hall opens one of the two doors you did not select that he knows is a losing door, selecting at random if both are losing doors. Monty asks you whether you would like to switch to this door. Suppose that the three doors in the puzzle are labeled 1, 2, and 3. Let W be the random variable whose value is the number of the winning door; assume that $p(W = k) = 1/3$ for $k = 1, 2, 3$. Let M denote the random variable whose value is the number of the door that Monty opens. Suppose you choose door i .
 - a) What is the probability that you will win the prize if the game ends before Monty asks you whether you want to change doors?
 - b) Find $p(M = j | W = k)$ for $j = 1, 2, 3$ and $k = 1, 2, 3$.
 - c) Use Bayes' Theorem to find $p(W = j | M = k)$ where i and j and k are distinct values.
 - d) Explain why the answer to part (c) tells you whether you should change doors when Monty gives you the chance to do so.
16. Ramesh can get to work three different ways: by bicycle, by car, or by bus. Because of commuter traffic, there is a 50% chance that he will be late when he drives his car. When he takes the bus, which uses a special lane reserved for buses, there is a 20% chance that he will be late. The probability that he is late when he rides his bicycle is only 5%. Ramesh arrives late one day. His boss wants to estimate the probability that he drove his car to work that day.
 - a) Suppose the boss assumes that there is a $1/3$ chance that Ramesh takes each of the three ways he can get to work. What estimate for the probability that Ramesh drove his car does the boss obtain from Bayes' Theorem under this assumption?
 - b) Suppose the boss knows that Ramesh drives 30% of the time, takes the bus only 10% of the time, and takes his bicycle 60% of the time. What estimate for the probability that Ramesh drove his car does the boss obtain from Bayes' Theorem using this information?
- *17. Prove Theorem 2, the extended form of Bayes' Theorem. That is, suppose that E is an event from a sample space S and that F_1, F_2, \dots, F_n are mutually exclusive events such that $\bigcup_{i=1}^n F_i = S$. Assume that $p(E) \neq 0$ and $p(F_i) \neq 0$ for $i = 1, 2, \dots, n$. Show that

$$p(F_j | E) = \frac{p(E | F_j)p(F_j)}{\sum_{i=1}^n p(E | F_i)p(F_i)}.$$

[Hint: Use the fact that $E = \bigcup_{i=1}^n (E \cap F_i)$.]
18. Suppose that a Bayesian spam filter is trained on a set of 500 spam messages and 200 messages that are not spam. The word "exciting" appears in 40 spam messages and in 25 messages that are not spam. Would an incoming message be rejected as spam if it contains the word "exciting" and the threshold for rejecting spam is 0.9?
19. Suppose that a Bayesian spam filter is trained on a set of 1000 spam messages and 400 messages that are not spam. The word "opportunity" appears in 175 spam messages and 20 messages that are not spam. Would an incoming message be rejected as spam if it contains the word "opportunity" and the threshold for rejecting a message is 0.9?
20. Would we reject a message as spam in Example 4
 - a) using just the fact that the word "undervalued" occurs in the message?
 - b) using just the fact that the word "stock" occurs in the message?
21. Suppose that a Bayesian spam filter is trained on a set of 10,000 spam messages and 5000 messages that are not spam. The word "enhancement" appears in 1500 spam messages and 20 messages that are not spam, while the word "herbal" appears in 800 spam messages and 200 messages that are not spam. Estimate the probability that a received message containing both the words "enhancement" and "herbal" is spam. Will the message be rejected as spam if the threshold for rejecting spam is 0.9?
22. Suppose that we have prior information concerning whether a random incoming message is spam. In particular, suppose that over a time period, we find that s spam messages arrive and h messages arrive that are not spam.
 - a) Use this information to estimate $p(S)$, the probability that an incoming message is spam, and $p(\bar{S})$, the probability an incoming message is not spam.
 - b) Use Bayes' Theorem and part (a) to estimate the probability that an incoming message containing the word w is spam, where $p(w)$ is the probability that w occurs in a spam message and $q(w)$ is the probability that w occurs in a message that is not spam.
23. Suppose that E_1 and E_2 are the events that an incoming mail message contains the words w_1 and w_2 , respectively. Assuming that E_1 and E_2 are independent events and that $E_1 | S$ and $E_2 | S$ are independent events, where S is the event that an incoming message is spam, and that we have no prior knowledge regarding whether or not the message is spam, show that

$$p(S | E_1 \cap E_2) = \frac{p(E_1 | S)p(E_2 | S)}{p(E_1 | S)p(E_2 | S) + p(E_1 | \bar{S})p(E_2 | \bar{S})}.$$

6.4 Expected Value and Variance

Introduction

The **expected value** of a random variable is the sum over all elements in a sample space of the product of the probability of the element and the value of the random variable at this element. Consequently, the expected value is a weighted average of the values of a random variable. The expected value of a random variable provides a central point for the distribution of values of this random variable. We can solve many problems using the notion of the expected value of a random variable, such as determining who has an advantage in gambling games and computing the average-case complexity of algorithms. Another useful measure of a random variable is its **variance**, which tells us how spread out the values of this random variable are. We can use the variance of a random variable to help us estimate the probability that a random variable takes values far removed from its expected value.

Expected Values



Many questions can be formulated in terms of the value we expect a random variable to take, or more precisely, the average value of a random variable when an experiment is performed a large number of times. Questions of this kind include: How many heads are expected to appear when a coin is flipped 100 times? What is the expected number of comparisons used to find an element in a list using a linear search? To study such questions we introduce the concept of the expected value of a random variable.

DEFINITION 1 The *expected value* (or *expectation*) of the random variable $X(s)$ on the sample space S is equal to

$$E(X) = \sum_{s \in S} p(s)X(s).$$

Note that when the sample space S has n elements $S = \{x_1, x_2, \dots, x_n\}$, $E(X) = \sum_{i=1}^n p(x_i)X(x_i)$.

Remark: When there are infinitely many elements of the sample space, the expectation is defined only when the infinite series in the definition is absolutely convergent. In particular, the expectation of a random variable on an infinite sample space is finite if it exists.

EXAMPLE 1 Expected Value of a Die Let X be the number that comes up when a die is rolled. What is the expected value of X ?

Solution: The random variable X takes the values 1, 2, 3, 4, 5, or 6, each with probability $1/6$. It follows that

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = \frac{7}{2}.$$

EXAMPLE 2 A fair coin is flipped three times. Let S be the sample space of the eight possible outcomes, and let X be the random variable that assigns to an outcome the number of heads in this outcome. What is the expected value of X ?



Solution: In Example 10 of Section 6.2 we listed the values of X for the eight possible outcomes when a coin is flipped three times. Because the coin is fair and the flips are independent, the probability of each outcome is $1/8$. Consequently,

$$\begin{aligned} E(X) &= \frac{1}{8}[X(HHH) + X(HHT) + X(HTH) + X(THH) + X(TTH) \\ &\quad + X(THT) + X(HTT) + X(TTT)] \\ &= \frac{1}{8}(3 + 2 + 2 + 2 + 1 + 1 + 1 + 0) = \frac{12}{8} \\ &= \frac{3}{2}. \end{aligned}$$

Consequently, the expected number of heads that come up when a fair coin is flipped three times is $3/2$. ◀

When an experiment has relatively few outcomes, we can compute the expected value of a random variable directly from its definition, as was done in Example 2. However, when an experiment has a large number of outcomes, it may be inconvenient to compute the expected value of a random variable directly from its definition. Instead, we can find the expected value of a random variable by grouping together all outcomes assigned the same value by the random variable, as Theorem 1 shows.

THEOREM 1 If X is a random variable and $p(X = r)$ is the probability that $X = r$, so that $p(X = r) = \sum_{s \in S, X(s)=r} p(s)$, then

$$E(X) = \sum_{r \in X(S)} p(X = r)r.$$

Proof: Suppose that X is a random variable with range $X(S)$, and let $p(X = r)$ be the probability that the random variable X takes the value r . Consequently, $p(X = r)$ is the sum of the probabilities of the outcomes s such that $X(s) = r$. It follows that

$$E(X) = \sum_{r \in X(S)} p(X = r)r. \quad \blacktriangleleft$$

Example 3 and the proof of Theorem 2 illustrate the use of this formula. In Example 3 we will find the expected value of the sum of the numbers that appear on two fair dice when they are rolled. In Theorem 2 we will find the expected value of the number of successes when n Bernoulli trials are performed.

EXAMPLE 3 What is the expected value of the sum of the numbers that appear when a pair of fair dice is rolled?

Solution: Let X be the random variable equal to the sum of the numbers that appear when a pair of dice is rolled. In Example 12 of Section 6.2 we listed the value of X for the 36 outcomes of

this experiment. The range of X is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. By Example 12 of Section 6.2 we see that

$$\begin{aligned}
 p(X = 2) &= p(X = 12) = 1/36, \\
 p(X = 3) &= p(X = 11) = 2/36 = 1/18, \\
 p(X = 4) &= p(X = 10) = 3/36 = 1/12, \\
 p(X = 5) &= p(X = 9) = 4/36 = 1/9, \\
 p(X = 6) &= p(X = 8) = 5/36, \\
 p(X = 7) &= 6/36 = 1/6.
 \end{aligned}$$

Substituting these values in the formula, we have

$$\begin{aligned}
 E(X) &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{1}{18} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{1}{6} \\
 &\quad + 8 \cdot \frac{5}{36} + 9 \cdot \frac{1}{9} + 10 \cdot \frac{1}{12} + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36} \\
 &= 7.
 \end{aligned}$$

THEOREM 2 The expected number of successes when n independent Bernoulli trials are performed, where p is the probability of success on each trial, is np .

Proof: Let X be the random variable equal to the number of successes in n trials. By Theorem 2 of Section 6.2 we see that $p(X = k) = C(n, k)p^k q^{n-k}$. Hence, we have

$$\begin{aligned}
 E(X) &= \sum_{k=1}^n kp(X = k) && \text{by Theorem 1} \\
 &= \sum_{k=1}^n kC(n, k)p^k q^{n-k} && \text{by Theorem 2 in Section 6.2} \\
 &= \sum_{k=1}^n nC(n-1, k-1)p^k q^{n-k} && \text{by Exercise 21 in Section 5.4} \\
 &= np \sum_{k=1}^n C(n-1, k-1)p^{k-1} q^{n-k} && \text{factoring } np \text{ from each term} \\
 &= np \sum_{j=0}^{n-1} C(n-1, j)p^j q^{n-1-j} && \text{shifting index of summation with } j = k-1 \\
 &= np(p+q)^{n-1} && \text{by the Binomial Theorem} \\
 &= np. && \text{because } p+q = 1
 \end{aligned}$$

This completes the proof because it shows that the expected number of successes in n independent Bernoulli trials is np . \blacktriangleleft

Next we will show that the hypothesis that the Bernoulli trials are independent in Theorem 2 is not necessary.

Linearity of Expectations

Theorem 3 tells us that expected values are linear. For example, the expected value of the sum of random variables is the sum of their expected values. We will find this property exceedingly useful.

THEOREM 3 If $X_i, i = 1, 2, \dots, n$ with n a positive integer, are random variables on S , and if a and b are real numbers, then

- (i) $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$
- (ii) $E(aX + b) = aE(X) + b$.

Proof: The first result for $n = 2$ follows directly from the definition of expected value, because

$$\begin{aligned} E(X_1 + X_2) &= \sum_{s \in S} p(s)(X_1(s) + X_2(s)) \\ &= \sum_{s \in S} p(s)X_1(s) + \sum_{s \in S} p(s)X_2(s) \\ &= E(X_1) + E(X_2). \end{aligned}$$

The case with n random variables follows easily using mathematical induction from the case of two random variables. Finally, note that

$$\begin{aligned} E(aX + b) &= \sum_{s \in S} p(s)(aX(s) + b) \\ &= a \sum_{s \in S} p(s)X(s) + b \sum_{s \in S} p(s) \\ &= aE(X) + b \text{ because } \sum_{s \in S} p(s) = 1. \end{aligned}$$

◀

Examples 4 and 5 illustrate how to use Theorem 3.

EXAMPLE 4 Use Theorem 3 to find the expected value of the sum of the numbers that appear when a pair of fair dice is rolled. (This was done in Example 3 without the benefit of this theorem.)

Solution: Let X_1 and X_2 be the random variables with $X_1((i, j)) = i$ and $X_2((i, j)) = j$, so that X_1 is the number appearing on the first die and X_2 is the number appearing on the second die. It is easy to see that $E(X_1) = E(X_2) = 7/2$ because both equal $(1 + 2 + 3 + 4 + 5 + 6)/6 = 21/6 = 7/2$. The sum of the two numbers that appear when the two dice are rolled is the sum $X_1 + X_2$. By Theorem 3, the expected value of the sum is $E(X_1 + X_2) = E(X_1) + E(X_2) = 7/2 + 7/2 = 7$. ◀

EXAMPLE 5 In the proof of Theorem 2 we found the expected value of the number of successes when n independent Bernoulli trials are performed, where p is the probability of success on each trial

by direct computation. Show how Theorem 3 can be used to derive this result where the Bernoulli trials are not necessarily independent.

Solution: Let X_i be the random variable with $X_i((t_1, t_2, \dots, t_n)) = 1$ if t_i is a success and $X_i((t_1, t_2, \dots, t_n)) = 0$ if t_i is a failure. The expected value of X_i is $E(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p$ for $i = 1, 2, \dots, n$. Let $X = X_1 + X_2 + \dots + X_n$, so that X counts the number of successes when these n Bernoulli trials are performed. Theorem 3, applied to the sum of n random variables, shows that $E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = np$. ◀

We can take advantage of the linearity of expectations to find the solutions of many seemingly difficult problems. The key step is to express a random variable whose expectation we wish to find as the sum of random variables whose expectations are easy to find. Examples 6 and 7 illustrate this technique.

EXAMPLE 6 Expected Value in the Hatcheck Problem A new employee checks the hats of n people at a restaurant, forgetting to put claim check numbers on the hats. When customers return for their hats, the checker gives them back hats chosen at random from the remaining hats. What is the expected number of hats that are returned correctly?

Solution: Let X be the random variable that equals the number of people who receive the correct hat from the checker. Let X_i be the random variable with $X_i = 1$ if the i th person receives the correct hat and $X_i = 0$ otherwise. It follows that

$$X = X_1 + X_2 + \dots + X_n.$$

Because it is equally likely that the checker returns any of the hats to this person, it follows that the probability that the i th person receives the correct hat is $1/n$. Consequently, by Theorem 1, for all i we have

$$E(X_i) = 1 \cdot p(X_i = 1) + 0 \cdot p(X_i = 0) = 1 \cdot 1/n + 0 = 1/n.$$

By the linearity of expectations (Theorem 3), it follows that

$$E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = n \cdot 1/n = 1.$$

Consequently, the average number of people who receive the correct hat is exactly 1. Note that this answer is independent of the number of people who have checked their hats! (We will find an explicit formula for the probability that no one receives the correct hat in Section 7.6.) ◀

EXAMPLE 7 Expected Number of Inversions in a Permutation The ordered pair (i, j) is called an **inversion** in a permutation of the first n positive integers if $i < j$ but j precedes i in the permutation. For instance, there are six inversions in the permutation 3, 5, 1, 4, 2; these inversions are

$$(1, 3), (1, 5), (2, 3), (2, 4), (2, 5), (4, 5).$$

To find the expected number of inversions in a random permutation of the first n positive integers, we let $I_{i,j}$ be the random variable on the set of all permutations of the first n positive integers with $I_{i,j} = 1$ if (i, j) is an inversion of the permutation and $I_{i,j} = 0$ otherwise. It follows that if X is the random variable equal to the number of inversions in the permutation, then

$$X = \sum_{1 \leq i < j \leq n} I_{i,j}.$$

Note that it is equally likely for i to precede j in a randomly chosen permutation as it is for j to precede i . (To see this, note that there are an equal number of permutations with each of these properties.) Consequently, for all pairs i and j we have

$$E(I_{i,j}) = 1 \cdot p(I_{i,j} = 1) + 0 \cdot p(I_{i,j} = 0) = 1 \cdot 1/2 + 0 = 1/2.$$

Because there are $\binom{n}{2}$ pairs i and j with $1 \leq i < j \leq n$ and by the linearity of expectations (Theorem 3), we have

$$E(X) = \sum_{1 \leq i < j \leq n} I_{i,j} = \binom{n}{2} \cdot \frac{1}{2} = \frac{n(n-1)}{4}.$$

It follows that there are an average of $n(n-1)/4$ inversions in a permutation of the first n positive integers. ◀

Average-Case Computational Complexity



Computing the average-case computational complexity of an algorithm can be interpreted as computing the expected value of a random variable. Let the sample space of an experiment be the set of possible inputs a_j , $j = 1, 2, \dots, n$, and let X be the random variable that assigns to a_j the number of operations used by the algorithm when given a_j as input. Based on our knowledge of the input, we assign a probability $p(a_j)$ to each possible input value a_j . Then, the average-case complexity of the algorithm is

$$E(X) = \sum_{j=1}^n p(a_j)X(a_j).$$

This is the expected value of X .

Finding the average-case computational complexity of an algorithm is usually much more difficult than finding its worst-case computational complexity, and often involves the use of sophisticated methods. However, there are some algorithms for which the analysis required to find the average-case computational complexity is not difficult. For instance, in Example 8 we will illustrate how to find the average-case computational complexity of the linear search algorithm under different assumptions concerning the probability that the element for which we search is an element of the list.

EXAMPLE 8 Average-Case Complexity of the Linear Search Algorithm We are given a real number x and a list of n distinct real numbers. The linear search algorithm, described in Section 3.1, locates x by successively comparing it to each element in the list, terminating when x is located or when all the elements have been examined and it has been determined that x is not in the list. What is the average-case computational complexity of the linear search algorithm if the probability that x is in the list is p and it is equally likely that x is any of the n elements in the list? (There are $n+1$ possible types of input: the n numbers in the list and a number not in the list, which we treat as a single input.)

Solution: In Example 4 of Section 3.3 we showed that $2i+1$ comparisons are used if x equals the i th element of the list and, in Example 2 of Section 3.3, that $2n+2$ comparisons are used if x is not in the list. The probability that x equals a_i , the i th element in the list, is p/n , and the

probability that x is not in the list is $q = 1 - p$. It follows that the average-case computational complexity of the linear search algorithm is

$$\begin{aligned}
 E &= \frac{3p}{n} + \frac{5p}{n} + \cdots + \frac{(2n+1)p}{n} + (2n+2)q \\
 &= \frac{p}{n}(3 + 5 + \cdots + (2n+1)) + (2n+2)q \\
 &= \frac{p}{n}((n+1)^2 - 1) + (2n+2)q \\
 &= p(n+2) + (2n+2)q.
 \end{aligned}$$

(The third equality follows from Example 2 of Section 4.1.) For instance, when x is guaranteed to be in the list, we have $p = 1$ (so the probability that $x = a_i$ is $1/n$ for each i) and $q = 0$. Then $E = n + 2$, as we showed in Example 4 in Section 3.3.

When p , the probability that x is in the list, is $1/2$, it follows that $q = 1 - p = 1/2$, so $E = (n+2)/2 + n + 1 = (3n+4)/2$. Similarly, if the probability that x is in the list is $3/4$, we have $p = 3/4$ and $q = 1/4$, so $E = 3(n+2)/4 + (n+1)/2 = (5n+8)/4$.

Finally, when x is guaranteed not to be in the list, we have $p = 0$ and $q = 1$. It follows that $E = 2n + 2$, which is not surprising because we have to search the entire list. ◀

Example 9 illustrates how the linearity of expectations can help us find the average-case complexity of a sorting algorithm, the insertion sort.

EXAMPLE 9 Average-Case Complexity of the Insertion Sort What is the average number of comparisons used by the insertion sort to sort n distinct elements?

Solution: We first suppose that X is the random variable equal to the number of comparisons used by the insertion sort (described in Section 3.1) to sort a list a_1, a_2, \dots, a_n of n distinct elements. Then $E(X)$ is the average number of comparisons used. (Recall that at step i for $i = 2, \dots, n$, the insertion sort inserts the i th element in the original list into the correct position in the sorted list of the first $i - 1$ elements of the original list.)

We let X_i be the random variable equal to the number of comparisons used to insert a_i into the proper position after the first $i - 1$ elements a_1, a_2, \dots, a_{i-1} have been sorted. Because

$$X = X_2 + X_3 + \cdots + X_n,$$

we can use the linearity of expectations to conclude that

$$E(X) = E(X_2 + X_3 + \cdots + X_n) = E(X_2) + E(X_3) + \cdots + E(X_n).$$

To find $E(X_i)$ for $i = 2, 3, \dots, n$, let $p_j(k)$ denote the probability that the largest of the first j elements in the list occurs at the k th position, that is, that $\max(a_1, a_2, \dots, a_j) = a_k$, where $1 \leq k \leq j$. Because the elements of the list are randomly distributed, it is equally likely for the largest element among the first j elements to occur at any position. Consequently, $p_j(k) = 1/j$. If $X_i(k)$ equals the number of comparisons used by the insertion sort if a_i is inserted into the k th position in the list once a_1, a_2, \dots, a_{i-1} have been sorted, it follows that $X_i(k) = k$. Because it

is possible that a_i is inserted in any of the first i positions, we find that

$$E(X_i) = \sum_{k=1}^i p_i(k) \cdot X_i(k) = \sum_{k=1}^i \frac{1}{i} \cdot k = \frac{1}{i} \cdot \sum_{k=1}^i k = \frac{1}{i} \cdot \frac{i(i+1)}{2} = \frac{i+1}{2}.$$

It follows that

$$\begin{aligned} E(X) &= \sum_{i=2}^n E(X_i) = \sum_{i=2}^n \frac{i+1}{2} = \frac{1}{2} \sum_{j=3}^{n+1} j \\ &= \frac{1}{2} \frac{(n+1)(n+2)}{2} - \frac{1}{2}(1+2) = \frac{n^2 + 3n - 4}{4}. \end{aligned}$$

To obtain the third of these equalities we shifted the index of summation, setting $j = i + 1$. To obtain the fourth equality, we used the formula $\sum_{k=1}^m k = m(m+1)/2$ (from Table 2 in Section 2.4) with $m = n + 1$, subtracting off the missing terms with $j = 1$ and $j = 2$. We conclude that the average number of comparisons used by the insertion sort to sort n elements equals $(n^2 + 3n - 4)/4$, which is $\Theta(n^2)$. ◀

The Geometric Distribution

We now turn our attention to a random variable with infinitely many possible outcomes.

EXAMPLE 10 Suppose that the probability that a coin comes up tails is p . This coin is flipped repeatedly until it comes up tails. What is the expected number of flips until this coin comes up tails?



Solution: We first note that the sample space consists of all sequences that begin with any number of heads, denoted by H , followed by a tail, denoted by T . Therefore, the sample space is the set $\{T, HT, HHT, HHHT, HHHHT, \dots\}$. Note that this is an infinite sample space. We can determine the probability of an element of the sample space by noting that the coin flips are independent and that the probability of a head is $1 - p$. Therefore, $p(T) = p$, $p(HT) = (1 - p)p$, $p(HHT) = (1 - p)^2 p$, and in general the probability that the coin is flipped n times before a tail comes up, that is, that $n - 1$ heads come up followed by a tail, is $(1 - p)^{n-1} p$. (Exercise 14 asks for a verification that the sum of the probabilities of the points in the sample space is 1.)

Now let X be the random variable equal to the number of flips in an element in the sample space. That is, $X(T) = 1$, $X(HT) = 2$, $X(HHT) = 3$, and so on. Note that $p(X = j) = (1 - p)^{j-1} p$. The expected number of flips until the coin comes up tails equals $E(X)$.

Using Theorem 1, we find that

$$E(X) = \sum_{j=1}^{\infty} j \cdot p(X = j) = \sum_{j=1}^{\infty} j(1 - p)^{j-1} p = p \sum_{j=1}^{\infty} j(1 - p)^{j-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

[The third equality in this chain follows from Table 2 in Section 2.4, which tells us that $\sum_{j=1}^{\infty} j(1 - p)^{j-1} = 1/(1 - (1 - p))^2 = 1/p^2$.] It follows that the expected number of times the coin is flipped until tails comes up is $1/p$. Note that when the coin is fair we have $p = 1/2$, so the expected number of flips until it comes up tails is $1/(1/2) = 2$. ◀

The random variable X that equals the number of flips expected before a coin comes up tails is an example of a random variable with a **geometric distribution**.

DEFINITION 2 A random variable X has a *geometric distribution with parameter p* if $p(X = k) = (1 - p)^{k-1}p$ for $k = 1, 2, 3, \dots$.

Geometric distributions arise in many applications because they are used to study the time required before a particular event happens, such as the time required before we find an object with a certain property, the number of attempts before an experiment succeeds, the number of times a product can be used before it fails, and so on.

When we computed the expected value of the number of flips required before a coin comes up tails, we proved Theorem 4.

THEOREM 4 If the random variable X has the geometric distribution with parameter p , then $E(X) = 1/p$.

Independent Random Variables

We have already discussed independent events. We will now define what it means for two random variables to be independent.

DEFINITION 3 The random variables X and Y on a sample space S are *independent* if

$$p(X(s) = r_1 \text{ and } Y(s) = r_2) = p(X(s) = r_1) \cdot p(Y(s) = r_2),$$

or in words, if the probability that $X(s) = r_1$ and $Y(s) = r_2$ equals the product of the probabilities that $X(s) = r_1$ and $Y(s) = r_2$, for all real numbers r_1 and r_2 .

EXAMPLE 11 Are the random variables X_1 and X_2 from Example 4 independent?



Solution: Let $S = \{1, 2, 3, 4, 5, 6\}$, and let $i \in S$ and $j \in S$. Because there are 36 possible outcomes when the pair of dice is rolled and each is equally likely, we have

$$p(X_1 = i \text{ and } X_2 = j) = 1/36.$$

Furthermore, $p(X_1 = i) = 1/6$ and $p(X_2 = j) = 1/6$, because the probability that i appears on the first die and the probability that j appears on the second die are both $1/6$. It follows that

$$p(X_1 = i \text{ and } X_2 = j) = \frac{1}{36} \quad \text{and} \quad p(X_1 = i)p(X_2 = j) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

so X_1 and X_2 are independent. ◀

EXAMPLE 12 Show that the random variables X_1 and $X = X_1 + X_2$, where X_1 and X_2 are as defined in Example 4, are not independent.

Solution: Note that $p(X_1 = 1 \text{ and } X = 12) = 0$, because $X_1 = 1$ means the number appearing on the first die is 1, which implies that the sum of the numbers appearing on the two dice cannot equal 12. On the other hand, $p(X_1 = 1) = 1/6$ and $p(X = 12) = 1/36$. Hence $p(X_1 = 1 \text{ and } X = 12) \neq p(X_1 = 1) \cdot p(X = 12)$. This counterexample shows that X_1 and X are not independent. ◀

The expected value of the product of two independent random variables is the product of their expected values, as Theorem 5 shows.

THEOREM 5 If X and Y are independent random variables on a space S , then $E(XY) = E(X)E(Y)$.

Proof: From the definition of expected value and because X and Y are independent random variables, it follows that

$$\begin{aligned}
 E(XY) &= \sum_{s \in S} X(s)Y(s)p(s) \\
 &= \sum_{r_1 \in X(S), r_2 \in Y(S)} r_1 r_2 \cdot p(X(s) = r_1 \text{ and } Y(s) = r_2) \\
 &= \sum_{r_1 \in X(S), r_2 \in Y(S)} r_1 r_2 \cdot p(X(s) = r_1) \cdot p(Y(s) = r_2) \\
 &= \left[\sum_{r_1 \in X(S)} r_1 p(X(s) = r_1) \right] \cdot \left[\sum_{r_2 \in Y(S)} r_2 p(Y(s) = r_2) \right] \\
 &= E(X)E(Y).
 \end{aligned}$$

This completes the proof. ◀

Note that when X and Y are random variables that are not independent, we cannot conclude that $E(XY) = E(X)E(Y)$, as Example 13 shows.

EXAMPLE 13 Let X and Y be random variables that count the number of heads and the number of tails when a coin is flipped twice. Because $p(X = 2) = 1/4$, $p(X = 1) = 1/2$, and $p(X = 0) = 1/4$, by Theorem 1 we have

$$E(X) = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} = 1.$$

A similar computation shows that $E(Y) = 1$. We note that $XY = 0$ when either two heads and no tails or two tails and no heads come up and that $XY = 1$ when one head and one tail come up. Hence,

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

It follows that

$$E(XY) \neq E(X)E(Y).$$

This does not contradict Theorem 5 because X and Y are not independent, as the reader should verify (see Exercise 16). ◀

Variance



The expected value of a random variable tells us its average value, but nothing about how widely its values are distributed. For example, if X and Y are the random variables on the set $S = \{1, 2, 3, 4, 5, 6\}$, with $X(s) = 0$ for all $s \in S$ and $Y(s) = -1$ if $s \in \{1, 2, 3\}$ and $Y(s) = 1$ if $s \in \{4, 5, 6\}$, then the expected values of X and Y are both zero. However, the random variable X never varies from 0, while the random variable Y always differs from 0 by 1. The variance of a random variable helps us characterize how widely a random variable is distributed.

DEFINITION 4 Let X be a random variable on a sample space S . The *variance* of X , denoted by $V(X)$, is

$$V(X) = \sum_{s \in S} (X(s) - E(X))^2 p(s).$$

The *standard deviation* of X , denoted $\sigma(X)$, is defined to be $\sqrt{V(X)}$.

Theorem 6 provides a useful simple expression for the variance of a random variable.

THEOREM 6 If X is a random variable on a sample space S , then $V(X) = E(X^2) - E(X)^2$.

Proof: Note that

$$\begin{aligned} V(X) &= \sum_{s \in S} (X(s) - E(X))^2 p(s) \\ &= \sum_{s \in S} X(s)^2 p(s) - 2E(X) \sum_{s \in S} X(s)p(s) + E(X)^2 \sum_{s \in S} p(s) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned}$$

We have used the fact that $\sum_{s \in S} p(s) = 1$ in the next-to-last step. ◀

EXAMPLE 14 What is the variance of the random variable X with $X(t) = 1$ if a Bernoulli trial is a success and $X(t) = 0$ if it is a failure, where p is the probability of success?



Solution: Because X takes only the values 0 and 1, it follows that $X^2(t) = X(t)$. Hence,

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p) = pq. \quad \blacktriangleleft$$

EXAMPLE 15 Variance of the Value of a Die What is the variance of the random variable X , where X is the number that comes up when a die is rolled?

Solution: We have $V(X) = E(X^2) - E(X)^2$. By Example 1 we know that $E(X) = 7/2$. To find $E(X^2)$ note that X^2 takes the values i^2 , $i = 1, 2, \dots, 6$, each with probability $1/6$. It follows that

$$E(X^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}.$$

We conclude that

$$V(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}. \quad \blacktriangleleft$$

EXAMPLE 16 What is the variance of the random variable $X((i, j)) = 2i$, where i is the number appearing on the first die and j is the number appearing on the second die, when two dice are rolled?

Solution: We will use Theorem 6 to find the variance of X . To do so, we need to find the expected values of X and X^2 . Note that because $p(X = k)$ is $1/6$ for $k = 2, 4, 6, 8, 10, 12$ and is 0 otherwise,

$$E(X) = (2 + 4 + 6 + 8 + 10 + 12)/6 = 7,$$

and

$$E(X^2) = (2^2 + 4^2 + 6^2 + 8^2 + 10^2 + 12^2)/6 = 182/3.$$

It follows from Theorem 6 that

$$V(X) = E(X^2) - E(X)^2 = 182/3 - 49 = 35/3. \quad \blacktriangleleft$$

Another useful fact about variances is that the variance of the sum of two independent random variables is the sum of their variances. This result is useful for computing the variance of the result of n independent Bernoulli trials, for instance.

THEOREM 7 If X and Y are two independent random variables on a sample space S , then $V(X + Y) = V(X) + V(Y)$. Furthermore, if X_i , $i = 1, 2, \dots, n$, with n a positive integer, are pairwise independent random variables on S , then $V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n)$.

Proof: From Theorem 6, we have

$$V(X + Y) = E((X + Y)^2) - E(X + Y)^2.$$

It follows that

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

Because X and Y are independent, by Theorem 5 we have $E(XY) = E(X)E(Y)$. It follows that

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

We leave the proof of the case with n pairwise independent random variables to the reader (Exercise 28). Such a proof can be constructed by generalizing the proof we have given for the case for two random variables. Note that it is not possible to use mathematical induction in a straightforward way to prove the general case (see Exercise 27). ◀

EXAMPLE 17 Find the variance and standard deviation of the random variable X whose value when two dice are rolled is $X((i, j)) = i + j$, where i is the number appearing on the first die and j is the number appearing on the second die.

Solution: Let X_1 and X_2 be the random variables defined by $X_1((i, j)) = i$ and $X_2((i, j)) = j$ for a roll of the dice. Then $X = X_1 + X_2$, and X_1 and X_2 are independent, as Example 11 showed. From Theorem 7 it follows that $V(X) = V(X_1) + V(X_2)$. A simple computation as in Example 16, together with Exercise 25 in the Supplementary Exercises at the end of the chapter, tells us that $V(X_1) = V(X_2) = 35/12$. Hence, $V(X) = 35/12 + 35/12 = 35/6$ and $\sigma(X) = \sqrt{35/6}$. ◀

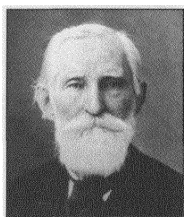
We will now find the variance of the random variable that counts the number of successes when n independent Bernoulli trials are carried out.

EXAMPLE 18 What is the variance of the number of successes when n independent Bernoulli trials are performed, where p is the probability of success on each trial?

Solution: Let X_i be the random variable with $X_i((t_1, t_2, \dots, t_n)) = 1$ if t_i is a success and $X_i((t_1, t_2, \dots, t_n)) = 0$ if t_i is a failure. Let $X = X_1 + X_2 + \dots + X_n$. Then X counts the number of successes in the n trials. From Theorem 7 it follows that $V(X) = V(X_1) + V(X_2) + \dots + V(X_n)$. Using Example 14 we have $V(X_i) = pq$ for $i = 1, 2, \dots, n$. It follows that $V(X) = npq$. ◀

Chebyshev's Inequality

How likely is it that a random variable takes a value far from its expected value? Theorem 8, called Chebyshev's Inequality, helps answer this question by providing an upper bound on the probability that the value of a random variable differs from the expected value of the random variable by more than a specified amount.



PAFNUTY LVOVICH CHEBYSHEV (1821–1894) Chebyshev was born into the gentry in Okatovo, Russia. His father was a retired army officer who had fought against Napoleon. In 1832 the family, with its nine children, moved to Moscow, where Pafnuty completed his high school education at home. He entered the Department of Physics and Mathematics at Moscow University. As a student, he developed a new method for approximating the roots of equations. He graduated from Moscow University in 1841 with a degree in mathematics, and he continued his studies, passing his master's exam in 1843 and completing his master's thesis in 1846.

Chebyshev was appointed in 1847 to a position as an assistant at the University of St. Petersburg. He wrote and defended a thesis in 1847. He became a professor at St. Petersburg in 1860, a position he held until 1882. His book on the theory of congruences written in 1849 was influential in the development of number theory. His

work on the distribution of prime numbers was seminal. He proved Bertrand's conjecture that for every integer $n > 3$, there is a prime between n and $2n - 2$. Chebyshev helped develop ideas that were later used to prove the Prime Number Theorem. Chebyshev's work on the approximation of functions using polynomials is used extensively when computers are used to find values of functions. Chebyshev was also interested in mechanics. He studied the conversion of rotary motion into rectilinear motion by mechanical coupling. The Chebyshev parallel motion is three linked bars approximating rectilinear motion.

THEOREM 8 CHEBYSHEV'S INEQUALITY Let X be a random variable on a sample space S with probability function p . If r is a positive real number, then

$$p(|X(s) - E(X)| \geq r) \leq V(X)/r^2.$$

Proof: Let A be the event

$$A = \{s \in S \mid |X(s) - E(X)| \geq r\}.$$

What we want to prove is that $p(A) \leq V(X)/r^2$. Note that

$$\begin{aligned} V(X) &= \sum_{s \in S} (X(s) - E(X))^2 p(s) \\ &= \sum_{s \in A} (X(s) - E(X))^2 p(s) + \sum_{s \notin A} (X(s) - E(X))^2 p(s). \end{aligned}$$

The second sum in this expression is nonnegative, because each of its summands is nonnegative. Also, because for each element s in A , $(X(s) - E(X))^2 \geq r^2$, the first sum in this expression is at least $\sum_{s \in A} r^2 p(s)$. Hence, $V(X) \geq \sum_{s \in A} r^2 p(s) = r^2 p(A)$. It follows that $V(X)/r^2 \geq p(A)$, so $p(A) \leq V(X)/r^2$, completing the proof. \blacktriangleleft

EXAMPLE 19 Deviation from the Mean when Counting Tails Suppose that X is the random variable that counts the number of tails when a fair coin is tossed n times. Note that X is the number of successes when n independent Bernoulli trials, each with probability of success $1/2$, are performed. It follows that $E(X) = n/2$ (by Theorem 2) and $V(X) = n/4$ (by Example 18). Applying Chebyshev's Inequality with $r = \sqrt{n}$ shows that

$$p(|X(s) - n/2| \geq \sqrt{n}) \leq (n/4)/(\sqrt{n})^2 = 1/4.$$

Consequently, the probability is no more than $1/4$ that the number of tails that come up when a fair coin is tossed n times deviates from the mean by more than \sqrt{n} . \blacktriangleleft

Chebyshev's Inequality, although applicable to any random variable, often fails to provide a practical estimate for the probability that the value of a random variable exceeds its mean by a large amount. This is illustrated by Example 20.

EXAMPLE 20 Let X be the random variable whose value is the number appearing when a fair die is rolled. We have $E(X) = 7/2$ (see Example 1) and $V(X) = 35/12$ (see Example 15). Because the only possible values of X are 1, 2, 3, 4, 5, and 6, X cannot take a value more than $5/2$ from its mean, $E(X) = 7/2$. Hence, $p(|X - 7/2| \geq r) = 0$ if $r > 5/2$. By Chebyshev's Inequality we know that $p(|X - 7/2| \geq r) \leq (35/12)/r^2$. For example, when $r = 3$, Chebyshev's Inequality tells us that $p(|X - 7/2| \geq 3) \leq (35/12)/9 = 35/108$, which is a poor estimate, because $p(|X - 7/2| \geq 3) = 0$. \blacktriangleleft

Exercises

1. What is the expected number of heads that come up when a fair coin is flipped five times?
2. What is the expected number of heads that come up when a fair coin is flipped 10 times?
3. What is the expected number of times a 6 appears when a fair die is rolled 10 times?
4. A coin is biased so that the probability a head comes up when it is flipped is 0.6. What is the expected number of heads that come up when it is flipped 10 times?

5. What is the expected sum of the numbers that appear on two dice, each biased so that a 3 comes up twice as often as each other number?
6. What is the expected value when a \$1 lottery ticket is bought in which the purchaser wins exactly \$10 million if the ticket contains the six winning numbers chosen from the set $\{1, 2, 3, \dots, 50\}$ and the purchaser wins nothing otherwise?
7. The final exam of a discrete mathematics course consists of 50 true/false questions, each worth two points, and 25 multiple-choice questions, each worth four points. The probability that Linda answers a true/false question correctly is 0.9, and the probability that she answers a multiple-choice question correctly is 0.8. What is her expected score on the final?
8. What is the expected sum of the numbers that appear when three fair dice are rolled?
9. Suppose that the probability that x is in a list of n distinct integers is $2/3$ and that it is equally likely that x equals any element in the list. Find the average number of comparisons used by the linear search algorithm to find x or to determine that it is not in the list.
10. Suppose that we flip a coin until either it comes up tails twice or we have flipped it six times. What is the expected number of times we flip the coin?
11. Suppose that we roll a die until a 6 comes up or we have rolled it 10 times. What is the expected number of times we roll the die?
12. Suppose that we roll a die until a 6 comes up.
 - a) What is the probability that we roll the die n times?
 - b) What is the expected number of times we roll the die?
13. Suppose that we roll a pair of dice until the sum of the numbers on the dice is seven. What is the expected number of times we roll the dice?
14. Show that the sum of the probabilities of a random variable with geometric distribution with parameter p , where $0 < p \leq 1$, equals 1.
15. Show that if the random variable X has the geometric distribution with parameter p , and j is a positive integer, then $p(X \geq j) = (1 - p)^{j-1}$.
16. Let X and Y be the random variables that count the number of heads and the number of tails that come up when two coins are flipped. Show that X and Y are not independent.
17. Estimate the expected number of integers with 1000 digits that need to be selected at random to find a prime, if the probability a number with 1000 digits is prime is approximately $1/2302$.
18. Suppose that X and Y are random variables and that X and Y are nonnegative for all points in a sample space S . Let Z be the random variable defined by $Z(s) = \max(X(s), Y(s))$ for all elements $s \in S$. Show that $E(Z) \leq E(X) + E(Y)$.
19. Let X be the number appearing on the first die when two dice are rolled and let Y be the sum of the numbers appearing on the two dice. Show that $E(X)E(Y) \neq E(XY)$.
20. Let A be an event. Then I_A , the **indicator random variable** of A , equals 1 if A occurs and equals 0 otherwise. Show that the expectation of the indicator random variable of A equals the probability of A , that is, $E(I_A) = p(A)$.
21. A **run** is a maximal sequence of successes in a sequence of Bernoulli trials. For example, in the sequence $S, S, S, F, S, S, F, F, S$, where S represents success and F represents failure, there are three runs consisting of three successes, two successes, and one success, respectively. Let R denote the random variable on the set of sequences of n independent Bernoulli trials that counts the number of runs in this sequence. Find $E(R)$. [Hint: Show that $R = \sum_{j=1}^n I_j$, where $I_j = 1$ if a run begins at the j th Bernoulli trial and $I_j = 0$ otherwise. Find $E(I_1)$ and then find $E(I_j)$, where $1 < j \leq n$.]
22. Let $X(s)$ be a random variable, where $X(s)$ is a nonnegative integer for all $s \in S$, and let A_k be the event that $X(s) \geq k$. Show that $E(X) = \sum_{k=1}^{\infty} p(A_k)$.
23. What is the variance of the number of heads that come up when a fair coin is flipped 10 times?
24. What is the variance of the number of times a 6 appears when a fair die is rolled 10 times?
25. Let X_n be the random variable that equals the number of tails minus the number of heads when n coins are flipped.
 - a) What is the expected value of X_n ?
 - b) What is the variance of X_n ?
26. Provide an example that shows that the variance of the sum of two random variables is not necessarily equal to the sum of their variances when the random variables are not independent.
27. Suppose that X_1 and X_2 are independent Bernoulli trials each with probability $1/2$, and that $X_3 = (X_1 + X_2) \bmod 2$.
 - a) Show that X_1 , X_2 , and X_3 are pairwise independent, but X_3 and $X_1 + X_2$ are not independent.
 - b) Show that $V(X_1 + X_2 + X_3) = V(X_1) + V(X_2) + V(X_3)$.
 - c) Explain why a proof by mathematical induction of Theorem 7 does not work by considering the random variables X_1 , X_2 , and X_3 .
- *28. Prove the general case of Theorem 7. That is, show that if X_1, X_2, \dots, X_n are pairwise independent random variables on a sample space S , where n is a positive integer, then $V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n)$. [Hint: Generalize the proof given in Theorem 7 for two random variables. Note that a proof using mathematical induction does not work; see Exercise 27.]
29. Use Chebyshev's Inequality to find an upper bound on the probability that the number of tails that come up when a fair coin is tossed n times deviates from the mean by more than $5\sqrt{n}$.
30. Use Chebyshev's Inequality to find an upper bound on the probability that the number of tails that come up when a

biased coin with probability of heads equal to 0.6 is tossed n times deviates from the mean by more than \sqrt{n} .

31. Let X be a random variable on a sample space S such that $X(s) \geq 0$ for all $s \in S$. Show that $p(X(s) \geq a) \leq E(X)/a$ for every positive real number a . This inequality is called **Markov's Inequality**.
32. Suppose that the number of cans of soda pop filled in a day at a bottling plant is a random variable with an expected value of 10,000 and a variance of 1000.
 - a) Use Markov's Inequality (Exercise 31) to obtain an upper bound on the probability that the plant will fill more than 11,000 cans on a particular day.
 - b) Use Chebyshev's Inequality to obtain a lower bound on the probability that the plant will fill between 9000 and 11,000 cans on a particular day.
33. Suppose that the number of tin cans recycled in a day at a recycling center is a random variable with an expected value of 50,000 and a variance of 10,000.
 - a) Use Markov's Inequality (Exercise 31) to find an upper bound on the probability that the center will recycle more than 55,000 cans on a particular day.
 - b) Use Chebyshev's Inequality to provide a lower bound on the probability that the center will recycle 40,000 to 60,000 cans on a certain day.
- *34. Suppose the probability that x is the i th element in a list of n distinct integers is $1/[n(n+1)]$. Find the average number of comparisons used by the linear search algorithm to find x or to determine that it is not in the list.
- *35. In this exercise we derive an estimate of the average-case complexity of the variant of the bubble sort algorithm that terminates once a pass has been made with no interchanges. Let X be the random variable on the set of permutations of a set of n distinct integers $\{a_1, a_2, \dots, a_n\}$ with $a_1 < a_2 < \dots < a_n$ such that $X(P)$ equals the number of comparisons used by the bubble sort to put these integers into increasing order.
 - a) Show that, under the assumption that the input is equally likely to be any of the $n!$ permutations of these integers, the average number of comparisons used by the bubble sort equals $E(X)$.
 - b) Use Example 5 in Section 3.3 to show that $E(X) \leq n(n-1)/2$.
 - c) Show that the sort makes at least one comparison for every inversion of two integers in the input.
 - d) Let $I(P)$ be the random variable that equals the number of inversions in the permutation P . Show that $E(X) \geq E(I)$.
 - e) Let $I_{j,k}$ be the random variable with $I_{j,k}(P) = 1$ if a_k precedes a_j in P and $I_{j,k} = 0$ otherwise. Show that $I(P) = \sum_k \sum_{j < k} I_{j,k}(P)$.
 - f) Show that $E(I) = \sum_k \sum_{j < k} E(I_{j,k})$.
 - g) Show that $E(I_{j,k}) = 1/2$. [Hint: Show that $E(I_{j,k})$ = probability that a_k precedes a_j in a permutation P . Then show it is equally likely for a_k to precede a_j as it is for a_j to precede a_k in a permutation.]
 - h) Use parts (f) and (g) to show that $E(I) = n(n-1)/4$.
- i) Conclude from parts (a), (b), and (h) that the average number of comparisons used to sort n integers is $\Theta(n^2)$.
- *36. In this exercise we find the average-case complexity of the quick sort algorithm, described in the preamble to Exercise 50 in Section 4.4, assuming a uniform distribution on the set of permutations.
 - a) Let X be the number of comparisons used by the quick sort algorithm to sort a list of n distinct integers. Show that the average number of comparisons used by the quick sort algorithm is $E(X)$ (where the sample space is the set of all $n!$ permutations of n integers).
 - b) Let $I_{j,k}$ denote the random variable that equals 1 if the j th smallest element and the k th smallest element of the initial list are ever compared as the quick sort algorithm sorts the list and equals 0 otherwise. Show that $X = \sum_{k=2}^n \sum_{j=1}^{k-1} I_{j,k}$.
 - c) Show that $E(X) = \sum_{k=2}^n \sum_{j=1}^{k-1} p$ (the j th smallest element and the k th smallest element are compared).
 - d) Show that p (the j th smallest element and the k th smallest element are compared), where $k > j$, equals $2/(k-j+1)$.
 - e) Use parts (c) and (d) to show that $E(X) = 2(n+1)(\sum_{i=2}^n 1/i) - 2(n-1)$.
 - f) Conclude from part (e) and the fact that $\sum_{j=1}^n 1/j \approx \ln n + \gamma$, where $\gamma = 0.57721\dots$ is Euler's constant, that the average number of comparisons used by the quick sort algorithm is $\Theta(n \log n)$.
- *37. What is the variance of the number of **fixed elements**, that is, elements left in the same position, of a randomly selected permutation of n elements? [Hint: Let X denote the number of fixed points of a random permutation. Write $X = X_1 + X_2 + \dots + X_n$, where $X_i = 1$ if the permutation fixes the i th element and $X_i = 0$ otherwise.]

The **covariance** of two random variables X and Y on a sample space S , denoted by $\text{Cov}(X, Y)$, is defined to be the expected value of the random variable $(X - E(X))(Y - E(Y))$. That is, $\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$.

38. Show that $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$, and use this result to conclude that $\text{Cov}(X, Y) = 0$ if X and Y are independent random variables.
39. Show that $V(X + Y) = V(X) + V(Y) + 2 \text{Cov}(X, Y)$.
40. Find $\text{Cov}(X, Y)$ if X and Y are the random variables with $X((i, j)) = 2i$ and $Y((i, j)) = i + j$, where i and j are the numbers that appear on the first and second of two dice when they are rolled.
41. When m balls are distributed into n bins uniformly at random, what is the probability that the first bin remains empty?
42. What is the expected number of balls that fall into the first bin when m balls are distributed into n bins uniformly at random?
43. What is the expected number of bins that remain empty when m balls are distributed into n bins uniformly at random?