Chapter 4

Conditional Probability

4.1 Discrete Conditional Probability

Conditional Probability

In this section we ask and answer the following question. Suppose we assign a distribution function to a sample space and then learn that an event E has occurred. How should we change the probabilities of the remaining events? We shall call the new probability for an event F the *conditional probability of F given E* and denote it by $P(F|E)$.

Example 4.1 An experiment consists of rolling a die once. Let X be the outcome. Let F be the event $\{X = 6\}$, and let E be the event $\{X > 4\}$. We assign the distribution function $m(\omega) = 1/6$ for $\omega = 1, 2, \ldots, 6$. Thus, $P(F) = 1/6$. Now suppose that the die is rolled and we are told that the event E has occurred. This leaves only two possible outcomes: 5 and 6. In the absence of any other information, we would still regard these outcomes to be equally likely, so the probability of F becomes $1/2$, making $P(F|E) = 1/2$. □

Example 4.2 In the Life Table (see Appendix C), one finds that in a population of 100,000 females, 89.835% can expect to live to age 60, while 57.062% can expect to live to age 80. Given that a woman is 60, what is the probability that she lives to age 80?

This is an example of a conditional probability. In this case, the original sample space can be thought of as a set of 100,000 females. The events E and F are the subsets of the sample space consisting of all women who live at least 60 years, and at least 80 years, respectively. We consider E to be the new sample space, and note that F is a subset of E. Thus, the size of E is 89,835, and the size of F is 57,062. So, the probability in question equals 57,062/89,835 = .6352. Thus, a woman who is 60 has a 63.52% chance of living to age 80. □

Example 4.3 Consider our voting example from Section 1.2: three candidates A, B, and C are running for office. We decided that A and B have an equal chance of winning and C is only 1/2 as likely to win as A. Let A be the event "A wins," B that "B wins," and C that "C wins." Hence, we assigned probabilities $P(A) = 2/5$, $P(B) = 2/5$, and $P(C) = 1/5$.

Suppose that before the election is held, A drops out of the race. As in Example 4.1, it would be natural to assign new probabilities to the events B and C which are proportional to the original probabilities. Thus, we would have $P(B|\ A) = 2/3$, and $P(C|\ A) = 1/3$. It is important to note that any time we assign probabilities to real-life events, the resulting distribution is only useful if we take into account all relevant information. In this example, we may have knowledge that most voters who favor A will vote for C if A is no longer in the race. This will clearly make the probability that C wins greater than the value of 1/3 that was assigned above. □

In these examples we assigned a distribution function and then were given new information that determined a new sample space, consisting of the outcomes that are still possible, and caused us to assign a new distribution function to this space.

We want to make formal the procedure carried out in these examples. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_r\}$ be the original sample space with distribution function $m(\omega_j)$ assigned. Suppose we learn that the event E has occurred. We want to assign a new distribution function $m(\omega_j|E)$ to Ω to reflect this fact. Clearly, if a sample point ω_j is not in E, we want $m(\omega_j|E) = 0$. Moreover, in the absence of information to the contrary, it is reasonable to assume that the probabilities for ω_k in E should have the same relative magnitudes that they had before we learned that E had occurred. For this we require that

$$m(\omega_k|E) = cm(\omega_k)$$

for all ω_k in E, with c some positive constant. But we must also have

$$\sum_E m(\omega_k|E) = c\sum_E m(\omega_k) = 1\ .$$

Thus,

$$c = \frac{1}{\sum_E m(\omega_k)} = \frac{1}{P(E)}\ .$$

(Note that this requires us to assume that $P(E) > 0$.) Thus, we will define

$$m(\omega_k|E) = \frac{m(\omega_k)}{P(E)}$$

for ω_k in E. We will call this new distribution the *conditional distribution* given E. For a general event F, this gives

$$P(F|E) = \sum_{F\cap E} m(\omega_k|E) = \sum_{F\cap E} \frac{m(\omega_k)}{P(E)} = \frac{P(F\cap E)}{P(E)}\ .$$

We call $P(F|E)$ the *conditional probability of F occurring given that E occurs,* and compute it using the formula

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

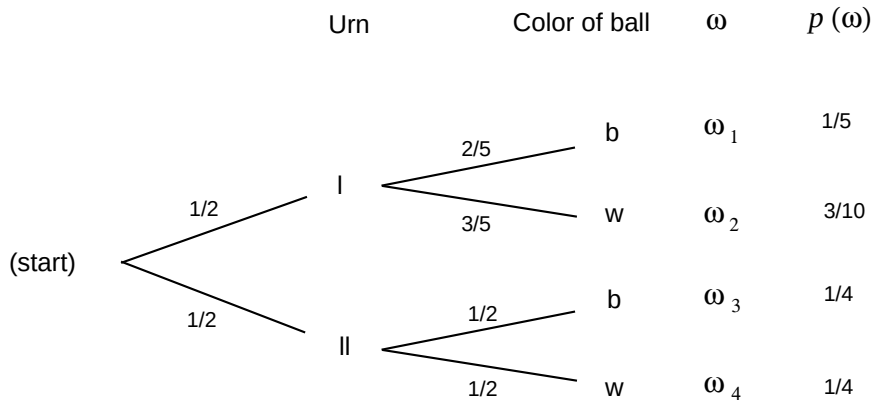

Figure 4.1: Tree diagram.

Example 4.4 (Example 4.1 continued) Let us return to the example of rolling a die. Recall that F is the event $X = 6$, and E is the event $X > 4$. Note that $E \cap F$ is the event F. So, the above formula gives

$$\begin{aligned} P(F|E) &= \frac{P(F \cap E)}{P(E)} \\ &= \frac{1/6}{1/3} \\ &= \frac{1}{2} , \end{aligned}$$

in agreement with the calculations performed earlier. □

Example 4.5 We have two urns, I and II. Urn I contains 2 black balls and 3 white balls. Urn II contains 1 black ball and 1 white ball. An urn is drawn at random and a ball is chosen at random from it. We can represent the sample space of this experiment as the paths through a tree as shown in Figure 4.1. The probabilities assigned to the paths are also shown.

Let B be the event "a black ball is drawn," and I the event "urn I is chosen." Then the branch weight 2/5, which is shown on one branch in the figure, can now be interpreted as the conditional probability $P(B|I)$.

Suppose we wish to calculate $P(I|B)$. Using the formula, we obtain

$$\begin{aligned} P(I|B) &= \frac{P(I \cap B)}{P(B)} \\ &= \frac{P(I \cap B)}{P(B \cap I) + P(B \cap II)} \\ &= \frac{1/5}{1/5 + 1/4} = \frac{4}{9} . \end{aligned}$$

□

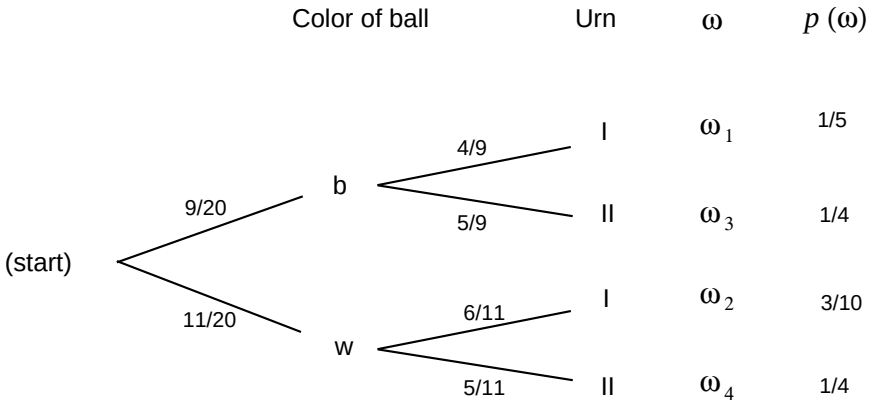

Figure 4.2: Reverse tree diagram.

Bayes Probabilities

Our original tree measure gave us the probabilities for drawing a ball of a given color, given the urn chosen. We have just calculated the *inverse probability* that a particular urn was chosen, given the color of the ball. Such an inverse probability is called a *Bayes probability* and may be obtained by a formula that we shall develop later. Bayes probabilities can also be obtained by simply constructing the tree measure for the two-stage experiment carried out in reverse order. We show this tree in Figure 4.2.

The paths through the reverse tree are in one-to-one correspondence with those in the forward tree, since they correspond to individual outcomes of the experiment, and so they are assigned the same probabilities. From the forward tree, we find that the probability of a black ball is

$$\frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{1}{2} = \frac{9}{20} \ .$$

The probabilities for the branches at the second level are found by simple division. For example, if x is the probability to be assigned to the top branch at the second level, we must have

$$\frac{9}{20} \cdot x = \frac{1}{5}$$

or $x = 4/9$. Thus, $P(I|B) = 4/9$, in agreement with our previous calculations. The reverse tree then displays all of the inverse, or Bayes, probabilities.

Example 4.6 We consider now a problem called the *Monty Hall* problem. This has long been a favorite problem but was revived by a letter from Craig Whitaker to Marilyn vos Savant for consideration in her column in *Parade Magazine*.[1] Craig wrote:

[1]Marilyn vos Savant, Ask Marilyn, *Parade Magazine*, 9 September; 2 December; 17 February 1990, reprinted in Marilyn vos Savant, *Ask Marilyn*, St. Martins, New York, 1992.

> Suppose you're on Monty Hall's *Let's Make a Deal!* You are given the choice of three doors, behind one door is a car, the others, goats. You pick a door, say 1, Monty opens another door, say 3, which has a goat. Monty says to you "Do you want to pick door 2?" Is it to your advantage to switch your choice of doors?

Marilyn gave a solution concluding that you should switch, and if you do, your probability of winning is 2/3. Several irate readers, some of whom identified themselves as having a PhD in mathematics, said that this is absurd since after Monty has ruled out one door there are only two possible doors and they should still each have the same probability 1/2 so there is no advantage to switching. Marilyn stuck to her solution and encouraged her readers to simulate the game and draw their own conclusions from this. We also encourage the reader to do this (see Exercise 11).

Other readers complained that Marilyn had not described the problem completely. In particular, the way in which certain decisions were made during a play of the game were not specified. This aspect of the problem will be discussed in Section 4.3. We will assume that the car was put behind a door by rolling a three-sided die which made all three choices equally likely. Monty knows where the car is, and always opens a door with a goat behind it. Finally, we assume that if Monty has a choice of doors (i.e., the contestant has picked the door with the car behind it), he chooses each door with probability 1/2. Marilyn clearly expected her readers to assume that the game was played in this manner.

As is the case with most apparent paradoxes, this one can be resolved through careful analysis. We begin by describing a simpler, related question. We say that a contestant is using the "stay" strategy if he picks a door, and, if offered a chance to switch to another door, declines to do so (i.e., he stays with his original choice). Similarly, we say that the contestant is using the "switch" strategy if he picks a door, and, if offered a chance to switch to another door, takes the offer. Now suppose that a contestant decides in advance to play the "stay" strategy. His only action in this case is to pick a door (and decline an invitation to switch, if one is offered). What is the probability that he wins a car? The same question can be asked about the "switch" strategy.

Using the "stay" strategy, a contestant will win the car with probability 1/3, since 1/3 of the time the door he picks will have the car behind it. On the other hand, if a contestant plays the "switch" strategy, then he will win whenever the door he originally picked does not have the car behind it, which happens 2/3 of the time.

This very simple analysis, though correct, does not quite solve the problem that Craig posed. Craig asked for the conditional probability that you win if you switch, given that you have chosen door 1 and that Monty has chosen door 3. To solve this problem, we set up the problem before getting this information and then compute the conditional probability given this information. This is a process that takes place in several stages; the car is put behind a door, the contestant picks a door, and finally Monty opens a door. Thus it is natural to analyze this using a tree measure. Here we make an additional assumption that if Monty has a choice

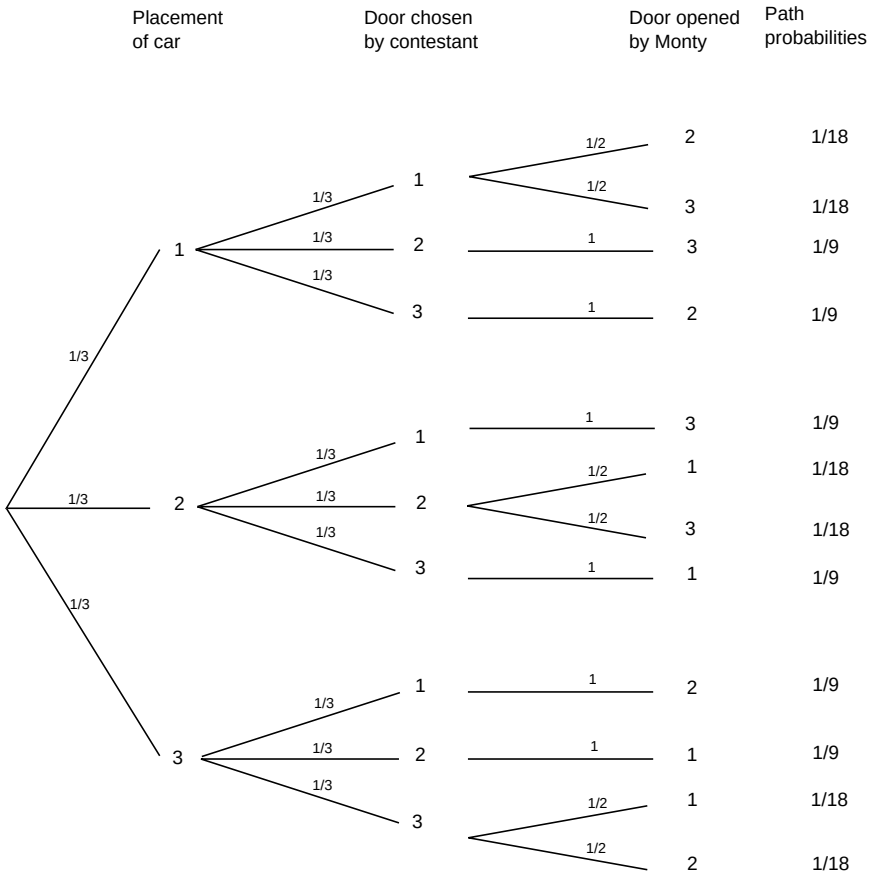

Figure 4.3: The Monty Hall problem.

of doors (i.e., the contestant has picked the door with the car behind it) then he picks each door with probability 1/2. The assumptions we have made determine the branch probabilities and these in turn determine the tree measure. The resulting tree and tree measure are shown in Figure 4.3. It is tempting to reduce the tree's size by making certain assumptions such as: "Without loss of generality, we will assume that the contestant always picks door 1." We have chosen not to make any such assumptions, in the interest of clarity.

Now the given information, namely that the contestant chose door 1 and Monty chose door 3, means only two paths through the tree are possible (see Figure 4.4). For one of these paths, the car is behind door 1 and for the other it is behind door 2. The path with the car behind door 2 is twice as likely as the one with the car behind door 1. Thus the conditional probability is 2/3 that the car is behind door 2 and 1/3 that it is behind door 1, so if you switch you have a 2/3 chance of winning the car, as Marilyn claimed.

At this point, the reader may think that the two problems above are the same, since they have the same answers. Recall that we assumed in the original problem

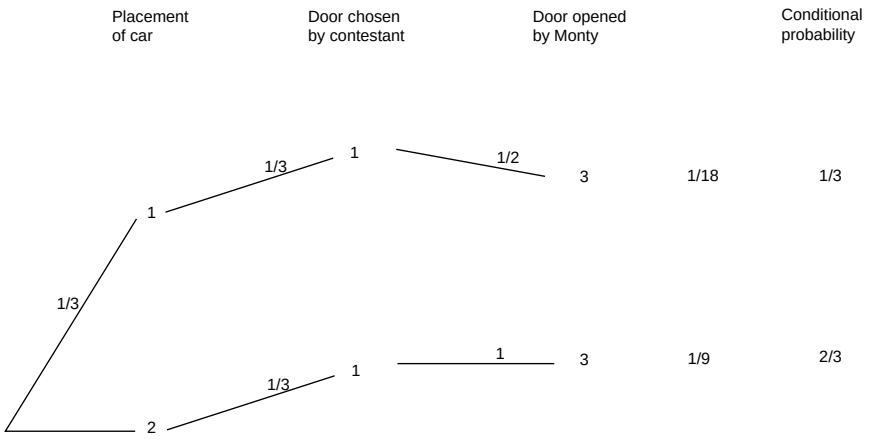

Figure 4.4: Conditional probabilities for the Monty Hall problem.

if the contestant chooses the door with the car, so that Monty has a choice of two doors, he chooses each of them with probability 1/2. Now suppose instead that in the case that he has a choice, he chooses the door with the larger number with probability 3/4. In the "switch" vs. "stay" problem, the probability of winning with the "switch" strategy is still 2/3. However, in the original problem, if the contestant switches, he wins with probability 4/7. The reader can check this by noting that the same two paths as before are the only two possible paths in the tree. The path leading to a win, if the contestant switches, has probability 1/3, while the path which leads to a loss, if the contestant switches, has probability 1/4. □

Independent Events

It often happens that the knowledge that a certain event E has occurred has no effect on the probability that some other event F has occurred, that is, that $P(F|E) = P(F)$. One would expect that in this case, the equation $P(E|F) = P(E)$ would also be true. In fact (see Exercise 1), each equation implies the other. If these equations are true, we might say the F is *independent* of E. For example, you would not expect the knowledge of the outcome of the first toss of a coin to change the probability that you would assign to the possible outcomes of the second toss, that is, you would not expect that the second toss depends on the first. This idea is formalized in the following definition of independent events.

Definition 4.1 Let E and F be two events. We say that they are *independent* if either 1) both events have positive probability and

$$P(E|F) = P(E) \text{ and } P(F|E) = P(F) \ ,$$

or 2) at least one of the events has probability 0. □

As noted above, if both $P(E)$ and $P(F)$ are positive, then each of the above equations imply the other, so that to see whether two events are independent, only one of these equations must be checked (see Exercise 1).

The following theorem provides another way to check for independence.

Theorem 4.1 Two events E and F are independent if and only if

$$P(E \cap F) = P(E)P(F) \ .$$

Proof. If either event has probability 0, then the two events are independent and the above equation is true, so the theorem is true in this case. Thus, we may assume that both events have positive probability in what follows. Assume that E and F are independent. Then $P(E|F) = P(E)$, and so

$$\begin{aligned} P(E \cap F) &= P(E|F)P(F) \\ &= P(E)P(F) \ . \end{aligned}$$

Assume next that $P(E \cap F) = P(E)P(F)$. Then

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

Also,

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

Therefore, E and F are independent. □

Example 4.7 Suppose that we have a coin which comes up heads with probability p, and tails with probability q. Now suppose that this coin is tossed twice. Using a frequency interpretation of probability, it is reasonable to assign to the outcome (H, H) the probability p^2, to the outcome (H, T) the probability pq, and so on. Let E be the event that heads turns up on the first toss and F the event that tails turns up on the second toss. We will now check that with the above probability assignments, these two events are independent, as expected. We have $P(E) = p^2 + pq = p$, $P(F) = pq + q^2 = q$. Finally $P(E \cap F) = pq$, so $P(E \cap F) = P(E)P(F)$. □

Example 4.8 It is often, but not always, intuitively clear when two events are independent. In Example 4.7, let A be the event "the first toss is a head" and B the event "the two outcomes are the same." Then

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P\{\text{HH}\}}{P\{\text{HH,HT}\}} = \frac{1/4}{1/2} = \frac{1}{2} = P(B).$$

Therefore, A and B are independent, but the result was not so obvious. □

Example 4.9 Finally, let us give an example of two events that are not independent. In Example 4.7, let I be the event "heads on the first toss" and J the event "two heads turn up." Then $P(I) = 1/2$ and $P(J) = 1/4$. The event $I \cap J$ is the event "heads on both tosses" and has probability $1/4$. Thus, I and J are not independent since $P(I)P(J) = 1/8 \neq P(I \cap J)$. □

We can extend the concept of independence to any finite set of events A_1, A_2, ..., A_n.

Definition 4.2 A set of events $\{A_1,\ A_2,\ \ldots,\ A_n\}$ is said to be *mutually independent* if for any subset $\{A_i,\ A_j, \ldots,\ A_m\}$ of these events we have

$$P(A_i \cap A_j \cap \cdots \cap A_m) = P(A_i)P(A_j)\cdots P(A_m),$$

or equivalently, if for any sequence $\bar{A}_1$, $\bar{A}_2$, ..., $\bar{A}_n$ with $\bar{A}_j = A_j$ or $\tilde{A}_j$,

$$P(\bar{A}_1 \cap \bar{A}_2 \cap \cdots \cap \bar{A}_n) = P(\bar{A}_1)P(\bar{A}_2)\cdots P(\bar{A}_n).$$

(For a proof of the equivalence in the case $n = 3$, see Exercise 33.) □

Using this terminology, it is a fact that any sequence $(\mathrm{S}, \mathrm{S}, \mathrm{F}, \mathrm{F}, \mathrm{S}, \ldots, \mathrm{S})$ of possible outcomes of a Bernoulli trials process forms a sequence of mutually independent events.

It is natural to ask: If all pairs of a set of events are independent, is the whole set mutually independent? The answer is *not necessarily,* and an example is given in Exercise 7.

It is important to note that the statement

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$$

does not imply that the events A_1, A_2, ..., A_n are mutually independent (see Exercise 8).

Joint Distribution Functions and Independence of Random Variables

It is frequently the case that when an experiment is performed, several different quantities concerning the outcomes are investigated.

Example 4.10 Suppose we toss a coin three times. The basic random variable $\bar{X}$ corresponding to this experiment has eight possible outcomes, which are the ordered triples consisting of H's and T's. We can also define the random variable X_i, for $i = 1, 2, 3$, to be the outcome of the ith toss. If the coin is fair, then we should assign the probability $1/8$ to each of the eight possible outcomes. Thus, the distribution functions of X_1, X_2, and X_3 are identical; in each case they are defined by $m(H) = m(T) = 1/2$. □

If we have several random variables $X_1, X_2, \ldots, X_n$ which correspond to a given experiment, then we can consider the joint random variable $\bar{X} = (X_1, X_2, \ldots, X_n)$ defined by taking an outcome ω of the experiment, and writing, as an n-tuple, the corresponding n outcomes for the random variables $X_1, X_2, \ldots, X_n$. Thus, if the random variable X_i has, as its set of possible outcomes the set R_i, then the set of possible outcomes of the joint random variable $\bar{X}$ is the Cartesian product of the R_i's, i.e., the set of all n-tuples of possible outcomes of the X_i's.

Example 4.11 (Example 4.10 continued) In the coin-tossing example above, let X_i denote the outcome of the ith toss. Then the joint random variable $\bar{X} = (X_1, X_2, X_3)$ has eight possible outcomes.

Suppose that we now define Y_i, for $i = 1, 2, 3$, as the number of heads which occur in the first i tosses. Then Y_i has $\{0, 1, \ldots, i\}$ as possible outcomes, so at first glance, the set of possible outcomes of the joint random variable $\bar{Y} = (Y_1, Y_2, Y_3)$ should be the set

$$\{(a_1, a_2, a_3) \ : \ 0 \leq a_1 \leq 1, 0 \leq a_2 \leq 2, 0 \leq a_3 \leq 3\} \ .$$

However, the outcome $(1, 0, 1)$ cannot occur, since we must have $a_1 \leq a_2 \leq a_3$. The solution to this problem is to define the probability of the outcome $(1, 0, 1)$ to be 0. In addition, we must have $a_{i+1} - a_i \leq 1$ for $i = 1, 2$.

We now illustrate the assignment of probabilities to the various outcomes for the joint random variables $\bar{X}$ and $\bar{Y}$. In the first case, each of the eight outcomes should be assigned the probability 1/8, since we are assuming that we have a fair coin. In the second case, since Y_i has $i + 1$ possible outcomes, the set of possible outcomes has size 24. Only eight of these 24 outcomes can actually occur, namely the ones satisfying $a_1 \leq a_2 \leq a_3$. Each of these outcomes corresponds to exactly one of the outcomes of the random variable $\bar{X}$, so it is natural to assign probability 1/8 to each of these. We assign probability 0 to the other 16 outcomes. In each case, the probability function is called a joint distribution function. □

We collect the above ideas in a definition.

Definition 4.3 Let $X_1, X_2, \ldots, X_n$ be random variables associated with an experiment. Suppose that the sample space (i.e., the set of possible outcomes) of X_i is the set R_i. Then the joint random variable $\bar{X} = (X_1, X_2, \ldots, X_n)$ is defined to be the random variable whose outcomes consist of ordered n-tuples of outcomes, with the ith coordinate lying in the set R_i. The sample space Ω of $\bar{X}$ is the Cartesian product of the R_i's:

$$\Omega = R_1 \times R_2 \times \cdots \times R_n \ .$$

The joint distribution function of $\bar{X}$ is the function which gives the probability of each of the outcomes of $\bar{X}$. □

Example 4.12 (Example 4.10 continued) We now consider the assignment of probabilities in the above example. In the case of the random variable $\bar{X}$, the probability of any outcome (a_1, a_2, a_3) is just the product of the probabilities $P(X_i = a_i)$,

	Not smoke	Smoke	Total
Not cancer	40	10	50
Cancer	7	3	10
Totals	47	13	60

Table 4.1: Smoking and cancer.

		S	
		0	1
C	0	40/60	10/60
	1	7/60	3/60

Table 4.2: Joint distribution.

for $i = 1, 2, 3$. However, in the case of $\bar{Y}$, the probability assigned to the outcome $(1, 1, 0)$ is not the product of the probabilities $P(Y_1 = 1)$, $P(Y_2 = 1)$, and $P(Y_3 = 0)$. The difference between these two situations is that the value of X_i does not affect the value of X_j, if $i \neq j$, while the values of Y_i and Y_j affect one another. For example, if $Y_1 = 1$, then Y_2 cannot equal 0. This prompts the next definition. □

Definition 4.4 The random variables X_1, X_2, ..., X_n are *mutually independent* if

$$\begin{aligned} &P(X_1 = r_1, X_2 = r_2, \ldots, X_n = r_n) \\ &\qquad = P(X_1 = r_1)P(X_2 = r_2) \cdots P(X_n = r_n) \end{aligned}$$

for any choice of $r_1, r_2, \ldots, r_n$. Thus, if $X_1,\ X_2, \ldots,\ X_n$ are mutually independent, then the joint distribution function of the random variable

$$\bar{X} = (X_1, X_2, \ldots, X_n)$$

is just the product of the individual distribution functions. When two random variables are mutually independent, we shall say more briefly that they are *independent.* □

Example 4.13 In a group of 60 people, the numbers who do or do not smoke and do or do not have cancer are reported as shown in Table 4.1. Let Ω be the sample space consisting of these 60 people. A person is chosen at random from the group. Let $C(\omega) = 1$ if this person has cancer and 0 if not, and $S(\omega) = 1$ if this person smokes and 0 if not. Then the joint distribution of $\{C, S\}$ is given in Table 4.2. For example $P(C = 0, S = 0) = 40/60$, $P(C = 0, S = 1) = 10/60$, and so forth. The distributions of the individual random variables are called *marginal distributions.* The marginal distributions of C and S are:

$$p_C = \begin{pmatrix} 0 & 1 \\ 50/60 & 10/60 \end{pmatrix},$$

$$p_S = \begin{pmatrix} 0 & 1 \\ 47/60 & 13/60 \end{pmatrix}.$$

The random variables S and C are not independent, since

$$\begin{aligned} P(C = 1, S = 1) &= \frac{3}{60} = .05 \ , \\ P(C = 1)P(S = 1) &= \frac{10}{60} \cdot \frac{13}{60} = .036 \ . \end{aligned}$$

Note that we would also see this from the fact that

$$\begin{aligned} P(C = 1|S = 1) &= \frac{3}{13} = .23 \ , \\ P(C = 1) &= \frac{1}{6} = .167 \ . \end{aligned}$$

□

Independent Trials Processes

The study of random variables proceeds by considering special classes of random variables. One such class that we shall study is the class of *independent trials.*

Definition 4.5 A sequence of random variables X_1, X_2, ..., X_n that are mutually independent and that have the same distribution is called a sequence of independent trials or an *independent trials process.*

Independent trials processes arise naturally in the following way. We have a single experiment with sample space $R = \{r_1, r_2, \ldots, r_s\}$ and a distribution function

$$m_X = \begin{pmatrix} r_1 & r_2 & \cdots & r_s \\ p_1 & p_2 & \cdots & p_s \end{pmatrix} \ .$$

We repeat this experiment n times. To describe this total experiment, we choose as sample space the space

$$\Omega = R \times R \times \cdots \times R,$$

consisting of all possible sequences $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ where the value of each ω_j is chosen from R. We assign a distribution function to be the *product distribution*

$$m(\omega) = m(\omega_1) \cdot \ \ldots \ \cdot m(\omega_n) \ ,$$

with $m(\omega_j) = p_k$ when $\omega_j = r_k$. Then we let X_j denote the jth coordinate of the outcome $(r_1, r_2, \ldots, r_n)$. The random variables X_1, ..., X_n form an independent trials process. □

Example 4.14 An experiment consists of rolling a die three times. Let X_i represent the outcome of the ith roll, for $i = 1, 2, 3$. The common distribution function is

$$m_i = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

The sample space is $R^3 = R \times R \times R$ with $R = \{1, 2, 3, 4, 5, 6\}$. If $\omega = (1, 3, 6)$, then $X_1(\omega) = 1$, $X_2(\omega) = 3$, and $X_3(\omega) = 6$ indicating that the first roll was a 1, the second was a 3, and the third was a 6. The probability assigned to any sample point is

$$m(\omega) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216} \ .$$

□

Example 4.15 Consider next a Bernoulli trials process with probability p for success on each experiment. Let $X_j(\omega) = 1$ if the jth outcome is success and $X_j(\omega) = 0$ if it is a failure. Then X_1, X_2, ..., X_n is an independent trials process. Each X_j has the same distribution function

$$m_j = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix},$$

where $q = 1 - p$.

If $S_n = X_1 + X_2 + \cdots + X_n$, then

$$P(S_n = j) = \binom{n}{j} p^j q^{n-j} \ ,$$

and S_n has, as distribution, the binomial distribution $b(n, p, j)$. □

Bayes' Formula

In our examples, we have considered conditional probabilities of the following form: Given the outcome of the second stage of a two-stage experiment, find the probability for an outcome at the first stage. We have remarked that these probabilities are called *Bayes probabilities.*

We return now to the calculation of more general Bayes probabilities. Suppose we have a set of events H_1, H_2, ..., H_m that are pairwise disjoint and such that the sample space Ω satisfies the equation

$$\Omega = H_1 \cup H_2 \cup \cdots \cup H_m \ .$$

We call these events *hypotheses.* We also have an event E that gives us some information about which hypothesis is correct. We call this event *evidence.*

Before we receive the evidence, then, we have a set of *prior probabilities* $P(H_1)$, $P(H_2)$, ..., $P(H_m)$ for the hypotheses. If we know the correct hypothesis, we know the probability for the evidence. That is, we know $P(E|H_i)$ for all i. We want to find the probabilities for the hypotheses given the evidence. That is, we want to find the conditional probabilities $P(H_i|E)$. These probabilities are called the *posterior probabilities.*

To find these probabilities, we write them in the form

$$P(H_i|E) = \frac{P(H_i \cap E)}{P(E)} \ . \tag{4.1}$$

Disease	Number having this disease	The results			
		+ +	+ −	− +	− −
d_1	3215	2110	301	704	100
d_2	2125	396	132	1187	410
d_3	4660	510	3568	73	509
Total	10000				

Table 4.3: Diseases data.

We can calculate the numerator from our given information by

$$P(H_i \cap E) = P(H_i)P(E|H_i) \ . \tag{4.2}$$

Since one and only one of the events H_1, H_2, ..., H_m can occur, we can write the probability of E as

$$P(E) = P(H_1 \cap E) + P(H_2 \cap E) + \cdots + P(H_m \cap E) \ .$$

Using Equation 4.2, the above expression can be seen to equal

$$P(H_1)P(E|H_1) + P(H_2)P(E|H_2) + \cdots + P(H_m)P(E|H_m) \ . \tag{4.3}$$

Using (4.1), (4.2), and (4.3) yields *Bayes' formula*:

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{\sum_{k=1}^{m} P(H_k)P(E|H_k)} \ .$$

Although this is a very famous formula, we will rarely use it. If the number of hypotheses is small, a simple tree measure calculation is easily carried out, as we have done in our examples. If the number of hypotheses is large, then we should use a computer.

Bayes probabilities are particularly appropriate for medical diagnosis. A doctor is anxious to know which of several diseases a patient might have. She collects evidence in the form of the outcomes of certain tests. From statistical studies the doctor can find the prior probabilities of the various diseases before the tests, and the probabilities for specific test outcomes, given a particular disease. What the doctor wants to know is the posterior probability for the particular disease, given the outcomes of the tests.

Example 4.16 A doctor is trying to decide if a patient has one of three diseases d_1, d_2, or d_3. Two tests are to be carried out, each of which results in a positive (+) or a negative (−) outcome. There are four possible test patterns ++, +−, −+, and −−. National records have indicated that, for 10,000 people having one of these three diseases, the distribution of diseases and test results are as in Table 4.3.

From this data, we can estimate the prior probabilities for each of the diseases and, given a particular disease, the probability of a particular test outcome. For example, the prior probability of disease d_1 may be estimated to be $3215/10{,}000 = .3215$. The probability of the test result +−, given disease d_1, may be estimated to be $301/3215 = .094$.

		d_1	d_2	d_3
+	+	.700	.131	.169
+	–	.075	.033	.892
–	+	.358	.604	.038
–	–	.098	.403	.499

Table 4.4: Posterior probabilities.

We can now use Bayes' formula to compute various posterior probabilities. The computer program **Bayes** computes these posterior probabilities. The results for this example are shown in Table 4.4.

We note from the outcomes that, when the test result is $++$, the disease d_1 has a significantly higher probability than the other two. When the outcome is $+-$, this is true for disease d_3. When the outcome is $-+$, this is true for disease d_2. Note that these statements might have been guessed by looking at the data. If the outcome is $--$, the most probable cause is d_3, but the probability that a patient has d_2 is only slightly smaller. If one looks at the data in this case, one can see that it might be hard to guess which of the two diseases d_2 and d_3 is more likely. □

Our final example shows that one has to be careful when the prior probabilities are small.

Example 4.17 A doctor gives a patient a test for a particular cancer. Before the results of the test, the only evidence the doctor has to go on is that 1 woman in 1000 has this cancer. Experience has shown that, in 99 percent of the cases in which cancer is present, the test is positive; and in 95 percent of the cases in which it is not present, it is negative. If the test turns out to be positive, what probability should the doctor assign to the event that cancer is present? An alternative form of this question is to ask for the relative frequencies of false positives and cancers.

We are given that prior(cancer) = .001 and prior(not cancer) = .999. We know also that $P(+|\text{cancer}) = .99$, $P(-|\text{cancer}) = .01$, $P(+|\text{not cancer}) = .05$, and $P(-|\text{not cancer}) = .95$. Using this data gives the result shown in Figure 4.5.

We see now that the probability of cancer given a positive test has only increased from .001 to .019. While this is nearly a twenty-fold increase, the probability that the patient has the cancer is still small. Stated in another way, among the positive results, 98.1 percent are false positives, and 1.9 percent are cancers. When a group of second-year medical students was asked this question, over half of the students incorrectly guessed the probability to be greater than .5. □

Historical Remarks

Conditional probability was used long before it was formally defined. Pascal and Fermat considered the *problem of points*: given that team A has won m games and team B has won n games, what is the probability that A will win the series? (See Exercises 40–42.) This is clearly a conditional probability problem.

In his book, Huygens gave a number of problems, one of which was:

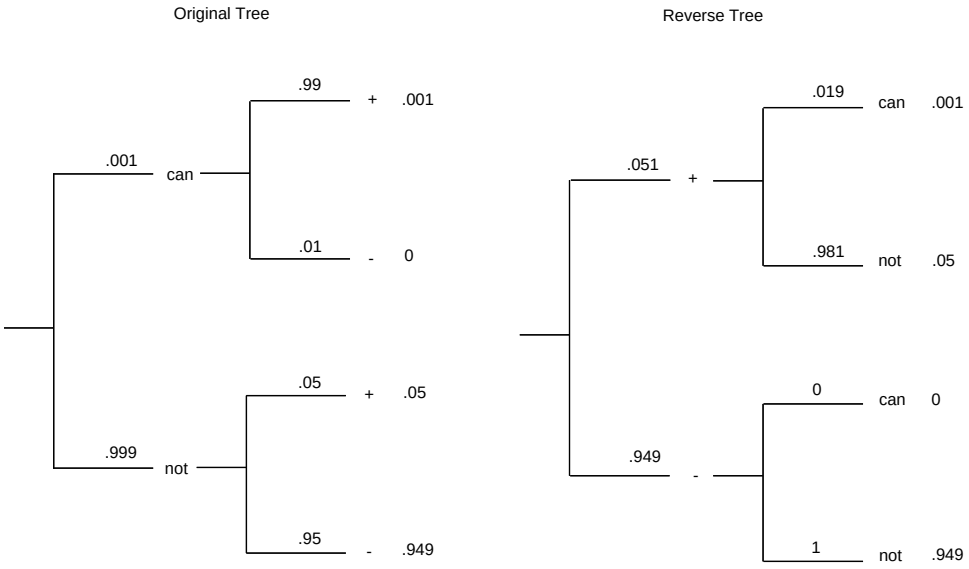

Figure 4.5: Forward and reverse tree diagrams.

> Three gamblers, A, B and C, take 12 balls of which 4 are white and 8 black. They play with the rules that the drawer is blindfolded, A is to draw first, then B and then C, the winner to be the one who first draws a white ball. What is the ratio of their chances?[2]

From his answer it is clear that Huygens meant that each ball is replaced after drawing. However, John Hudde, the mayor of Amsterdam, assumed that he meant to sample without replacement and corresponded with Huygens about the difference in their answers. Hacking remarks that "Neither party can understand what the other is doing."[3]

By the time of de Moivre's book, *The Doctrine of Chances,* these distinctions were well understood. De Moivre defined independence and dependence as follows:

> Two Events are independent, when they have no connexion one with the other, and that the happening of one neither forwards nor obstructs the happening of the other.
>
> Two Events are dependent, when they are so connected together as that the Probability of either's happening is altered by the happening of the other.[4]

De Moivre used sampling with and without replacement to illustrate that the probability that two independent events both happen is the product of their probabilities, and for dependent events that:

[2]Quoted in F. N. David, *Games, Gods and Gambling* (London: Griffin, 1962), p. 119.

[3]I. Hacking, *The Emergence of Probability* (Cambridge: Cambridge University Press, 1975), p. 99.

[4]A. de Moivre, *The Doctrine of Chances,* 3rd ed. (New York: Chelsea, 1967), p. 6.

> The Probability of the happening of two Events dependent, is the product of the Probability of the happening of one of them, by the Probability which the other will have of happening, when the first is considered as having happened; and the same Rule will extend to the happening of as many Events as may be assigned.[5]

The formula that we call Bayes' formula, and the idea of computing the probability of a hypothesis given evidence, originated in a famous essay of Thomas Bayes. Bayes was an ordained minister in Tunbridge Wells near London. His mathematical interests led him to be elected to the Royal Society in 1742, but none of his results were published within his lifetime. The work upon which his fame rests, "An Essay Toward Solving a Problem in the Doctrine of Chances," was published in 1763, three years after his death.[6] Bayes reviewed some of the basic concepts of probability and then considered a new kind of inverse probability problem requiring the use of conditional probability.

Bernoulli, in his study of processes that we now call Bernoulli trials, had proven his famous law of large numbers which we will study in Chapter 8. This theorem assured the experimenter that if he knew the probability p for success, he could predict that the proportion of successes would approach this value as he increased the number of experiments. Bernoulli himself realized that in most interesting cases you do not know the value of p and saw his theorem as an important step in showing that you could determine p by experimentation.

To study this problem further, Bayes started by assuming that the probability p for success is itself determined by a random experiment. He assumed in fact that this experiment was such that this value for p is equally likely to be any value between 0 and 1. Without knowing this value we carry out n experiments and observe m successes. Bayes proposed the problem of finding the conditional probability that the unknown probability p lies between a and b. He obtained the answer:

$$P(a \leq p < b | m \text{ successes in } n \text{ trials}) = \frac{\int_a^b x^m (1-x)^{n-m}\, dx}{\int_0^1 x^m (1-x)^{n-m}\, dx}\ .$$

We shall see in the next section how this result is obtained. Bayes clearly wanted to show that the conditional distribution function, given the outcomes of more and more experiments, becomes concentrated around the true value of p. Thus, Bayes was trying to solve an *inverse problem.* The computation of the integrals was too difficult for exact solution except for small values of j and n, and so Bayes tried approximate methods. His methods were not very satisfactory and it has been suggested that this discouraged him from publishing his results.

However, his paper was the first in a series of important studies carried out by Laplace, Gauss, and other great mathematicians to solve inverse problems. They studied this problem in terms of errors in measurements in astronomy. If an astronomer were to know the true value of a distance and the nature of the random

[5] ibid, p. 7.

[6] T. Bayes, "An Essay Toward Solving a Problem in the Doctrine of Chances," *Phil. Trans. Royal Soc. London,* vol. 53 (1763), pp. 370–418.

errors caused by his measuring device he could predict the probabilistic nature of his measurements. In fact, however, he is presented with the inverse problem of knowing the nature of the random errors, and the values of the measurements, and wanting to make inferences about the unknown true value.

As Maistrov remarks, the formula that we have called Bayes' formula does not appear in his essay. Laplace gave it this name when he studied these inverse problems.[7] The computation of inverse probabilities is fundamental to statistics and has led to an important branch of statistics called Bayesian analysis, assuring Bayes eternal fame for his brief essay.

Exercises

1 Assume that E and F are two events with positive probabilities. Show that if $P(E|F) = P(E)$, then $P(F|E) = P(F)$.

2 A coin is tossed three times. What is the probability that exactly two heads occur, given that

(a) the first outcome was a head?

(b) the first outcome was a tail?

(c) the first two outcomes were heads?

(d) the first two outcomes were tails?

(e) the first outcome was a head and the third outcome was a head?

3 A die is rolled twice. What is the probability that the sum of the faces is greater than 7, given that

(a) the first outcome was a 4?

(b) the first outcome was greater than 3?

(c) the first outcome was a 1?

(d) the first outcome was less than 5?

4 A card is drawn at random from a deck of cards. What is the probability that

(a) it is a heart, given that it is red?

(b) it is higher than a 10, given that it is a heart? (Interpret J, Q, K, A as 11, 12, 13, 14.)

(c) it is a jack, given that it is red?

5 A coin is tossed three times. Consider the following events
A: Heads on the first toss.
B: Tails on the second.
C: Heads on the third toss.
D: All three outcomes the same (HHH or TTT).
E: Exactly one head turns up.

[7] L. E. Maistrov, *Probability Theory: A Historical Sketch,* trans. and ed. Samual Kotz (New York: Academic Press, 1974), p. 100.

(a) Which of the following pairs of these events are independent?
(1) A, B
(2) A, D
(3) A, E
(4) D, E

(b) Which of the following triples of these events are independent?
(1) A, B, C
(2) A, B, D
(3) C, D, E

6 From a deck of five cards numbered 2, 4, 6, 8, and 10, respectively, a card is drawn at random and replaced. This is done three times. What is the probability that the card numbered 2 was drawn exactly two times, given that the sum of the numbers on the three draws is 12?

7 A coin is tossed twice. Consider the following events.
A: Heads on the first toss.
B: Heads on the second toss.
C: The two tosses come out the same.

(a) Show that A, B, C are pairwise independent but not independent.

(b) Show that C is independent of A and B but not of $A \cap B$.

8 Let $\Omega = \{a, b, c, d, e, f\}$. Assume that $m(a) = m(b) = 1/8$ and $m(c) = m(d) = m(e) = m(f) = 3/16$. Let A, B, and C be the events $A = \{d, e, a\}$, $B = \{c, e, a\}$, $C = \{c, d, a\}$. Show that $P(A \cap B \cap C) = P(A)P(B)P(C)$ but no two of these events are independent.

9 What is the probability that a family of two children has

(a) two boys given that it has at least one boy?

(b) two boys given that the first child is a boy?

10 In Example 4.2, we used the Life Table (see Appendix C) to compute a conditional probability. The number 93,753 in the table, corresponding to 40-year-old males, means that of all the males born in the United States in 1950, 93.753% were alive in 1990. Is it reasonable to use this as an estimate for the probability of a male, born this year, surviving to age 40?

11 Simulate the Monty Hall problem. Carefully state any assumptions that you have made when writing the program. Which version of the problem do you think that you are simulating?

12 In Example 4.17, how large must the prior probability of cancer be to give a posterior probability of .5 for cancer given a positive test?

13 Two cards are drawn from a bridge deck. What is the probability that the second card drawn is red?

14 If $P(\tilde{B}) = 1/4$ and $P(A|B) = 1/2$, what is $P(A \cap B)$?

15 (a) What is the probability that your bridge partner has exactly two aces, given that she has at least one ace?

(b) What is the probability that your bridge partner has exactly two aces, given that she has the ace of spades?

16 Prove that for any three events A, B, C, each having positive probability, and with the property that $P(A \cap B) > 0$,

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) \ .$$

17 Prove that if A and B are independent so are

(a) A and $\tilde{B}$.

(b) $\tilde{A}$ and $\tilde{B}$.

18 A doctor assumes that a patient has one of three diseases d_1, d_2, or d_3. Before any test, he assumes an equal probability for each disease. He carries out a test that will be positive with probability .8 if the patient has d_1, .6 if he has disease d_2, and .4 if he has disease d_3. Given that the outcome of the test was positive, what probabilities should the doctor now assign to the three possible diseases?

19 In a poker hand, John has a very strong hand and bets 5 dollars. The probability that Mary has a better hand is .04. If Mary had a better hand she would raise with probability .9, but with a poorer hand she would only raise with probability .1. If Mary raises, what is the probability that she has a better hand than John does?

20 The Polya urn model for contagion is as follows: We start with an urn which contains one white ball and one black ball. At each second we choose a ball at random from the urn and replace this ball and add one more of the color chosen. Write a program to simulate this model, and see if you can make any predictions about the proportion of white balls in the urn after a large number of draws. Is there a tendency to have a large fraction of balls of the same color in the long run?

21 It is desired to find the probability that in a bridge deal each player receives an ace. A student argues as follows. It does not matter where the first ace goes. The second ace must go to one of the other three players and this occurs with probability 3/4. Then the next must go to one of two, an event of probability 1/2, and finally the last ace must go to the player who does not have an ace. This occurs with probability 1/4. The probability that all these events occur is the product $(3/4)(1/2)(1/4) = 3/32$. Is this argument correct?

22 One coin in a collection of 65 has two heads. The rest are fair. If a coin, chosen at random from the lot and then tossed, turns up heads 6 times in a row, what is the probability that it is the two-headed coin?

23 You are given two urns and fifty balls. Half of the balls are white and half are black. You are asked to distribute the balls in the urns with no restriction placed on the number of either type in an urn. How should you distribute the balls in the urns to maximize the probability of obtaining a white ball if an urn is chosen at random and a ball drawn out at random? Justify your answer.

24 A fair coin is thrown n times. Show that the conditional probability of a head on any specified trial, given a total of k heads over the n trials, is k/n $(k > 0)$.

25 (Johnsonbough[8]) A coin with probability p for heads is tossed n times. Let E be the event "a head is obtained on the first toss' and F_k the event 'exactly k heads are obtained." For which pairs (n, k) are E and F_k independent?

26 Suppose that A and B are events such that $P(A|B) = P(B|A)$ and $P(A \cup B) = 1$ and $P(A \cap B) > 0$. Prove that $P(A) > 1/2$.

27 (Chung[9]) In London, half of the days have some rain. The weather forecaster is correct 2/3 of the time, i.e., the probability that it rains, given that she has predicted rain, and the probability that it does not rain, given that she has predicted that it won't rain, are both equal to 2/3. When rain is forecast, Mr. Pickwick takes his umbrella. When rain is not forecast, he takes it with probability 1/3. Find

(a) the probability that Pickwick has no umbrella, given that it rains.

(b) the probability that he brings his umbrella, given that it doesn't rain.

28 Probability theory was used in a famous court case: *People v. Collins.*[10] In this case a purse was snatched from an elderly person in a Los Angeles suburb. A couple seen running from the scene were described as a black man with a beard and a mustache and a blond girl with hair in a ponytail. Witnesses said they drove off in a partly yellow car. Malcolm and Janet Collins were arrested. He was black and though clean shaven when arrested had evidence of recently having had a beard and a mustache. She was blond and usually wore her hair in a ponytail. They drove a partly yellow Lincoln. The prosecution called a professor of mathematics as a witness who suggested that a conservative set of probabilities for the characteristics noted by the witnesses would be as shown in Table 4.5.

The prosecution then argued that the probability that all of these characteristics are met by a randomly chosen couple is the product of the probabilities or 1/12,000,000, which is very small. He claimed this was proof beyond a reasonable doubt that the defendants were guilty. The jury agreed and handed down a verdict of guilty of second-degree robbery.

[8]R. Johnsonbough, "Problem #103," *Two Year College Math Journal,* vol. 8 (1977), p. 292.

[9]K. L. Chung, *Elementary Probability Theory With Stochastic Processes, 3rd ed.* (New York: Springer-Verlag, 1979), p. 152.

[10]M. W. Gray, "Statistics and the Law," *Mathematics Magazine,* vol. 56 (1983), pp. 67–81.

man with mustache	1/4
girl with blond hair	1/3
girl with ponytail	1/10
black man with beard	1/10
interracial couple in a car	1/1000
partly yellow car	1/10

Table 4.5: Collins case probabilities.

If you were the lawyer for the Collins couple how would you have countered the above argument? (The appeal of this case is discussed in Exercise 5.1.34.)

29 A student is applying to Harvard and Dartmouth. He estimates that he has a probability of .5 of being accepted at Dartmouth and .3 of being accepted at Harvard. He further estimates the probability that he will be accepted by both is .2. What is the probability that he is accepted by Dartmouth if he is accepted by Harvard? Is the event "accepted at Harvard" independent of the event "accepted at Dartmouth"?

30 Luxco, a wholesale lightbulb manufacturer, has two factories. Factory A sells bulbs in lots that consists of 1000 regular and 2000 *softglow* bulbs each. Random sampling has shown that on the average there tend to be about 2 bad regular bulbs and 11 bad softglow bulbs per lot. At factory B the lot size is reversed—there are 2000 regular and 1000 softglow per lot—and there tend to be 5 bad regular and 6 bad softglow bulbs per lot.

The manager of factory A asserts, "We're obviously the better producer; our bad bulb rates are .2 percent and .55 percent compared to B's .25 percent and .6 percent. We're better at both regular and softglow bulbs by half of a tenth of a percent each."

"Au contraire," counters the manager of B, "each of our 3000 bulb lots contains only 11 bad bulbs, while A's 3000 bulb lots contain 13. So our .37 percent bad bulb rate beats their .43 percent."

Who is right?

31 Using the Life Table for 1981 given in Appendix C, find the probability that a male of age 60 in 1981 lives to age 80. Find the same probability for a female.

32 (a) There has been a blizzard and Helen is trying to drive from Woodstock to Tunbridge, which are connected like the top graph in Figure 4.6. Here p and q are the probabilities that the two roads are passable. What is the probability that Helen can get from Woodstock to Tunbridge?

(b) Now suppose that Woodstock and Tunbridge are connected like the middle graph in Figure 4.6. What now is the probability that she can get from W to T? Note that if we think of the roads as being components of a system, then in (a) and (b) we have computed the *reliability* of a system whose components are (a) *in series* and (b) *in parallel.*

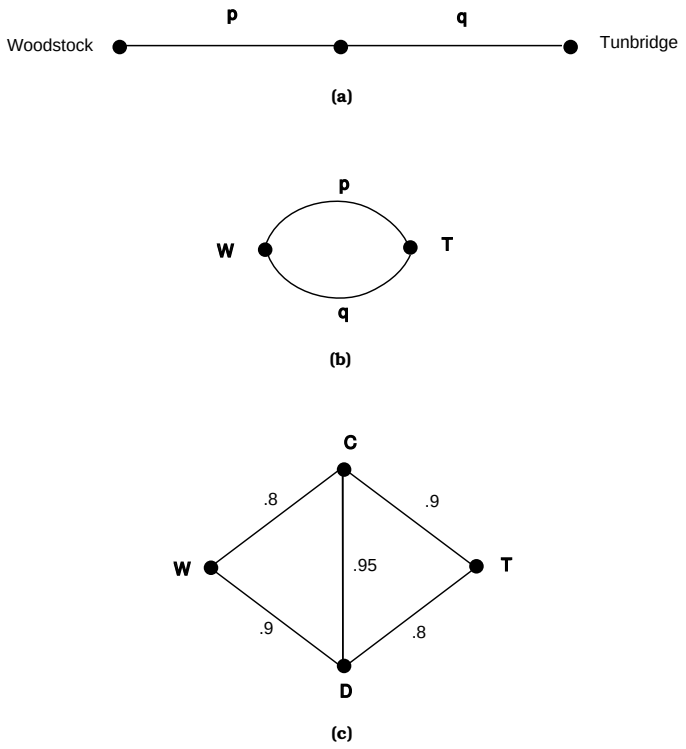

Figure 4.6: From Woodstock to Tunbridge.

(c) Now suppose W and T are connected like the bottom graph in Figure 4.6. Find the probability of Helen's getting from W to T. *Hint*: If the road from C to D is impassable, it might as well not be there at all; if it is passable, then figure out how to use part (b) twice.

33 Let A_1, A_2, and A_3 be events, and let B_i represent either A_i or its complement $\tilde{A}_i$. Then there are eight possible choices for the triple (B_1, B_2, B_3). Prove that the events A_1, A_2, A_3 are independent if and only if

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2)P(B_3) \ ,$$

for all eight of the possible choices for the triple (B_1, B_2, B_3).

34 Four women, A, B, C, and D, check their hats, and the hats are returned in a random manner. Let Ω be the set of all possible permutations of A, B, C, D. Let $X_j = 1$ if the jth woman gets her own hat back and 0 otherwise. What is the distribution of X_j? Are the X_i's mutually independent?

35 A box has numbers from 1 to 10. A number is drawn at random. Let X_1 be the number drawn. This number is replaced, and the ten numbers mixed. A second number X_2 is drawn. Find the distributions of X_1 and X_2. Are X_1 and X_2 independent? Answer the same questions if the first number is not replaced before the second is drawn.

		Y			
		-1	0	1	2
X	-1	0	1/36	1/6	1/12
	0	1/18	0	1/18	0
	1	0	1/36	1/6	1/12
	2	1/12	0	1/12	1/6

Table 4.6: Joint distribution.

36 A die is thrown twice. Let X_1 and X_2 denote the outcomes. Define $X = \min(X_1, X_2)$. Find the distribution of X.

***37** Given that $P(X = a) = r$, $P(\max(X, Y) = a) = s$, and $P(\min(X, Y) = a) = t$, show that you can determine $u = P(Y = a)$ in terms of r, s, and t.

38 A fair coin is tossed three times. Let X be the number of heads that turn up on the first two tosses and Y the number of heads that turn up on the third toss. Give the distribution of

(a) the random variables X and Y.

(b) the random variable $Z = X + Y$.

(c) the random variable $W = X - Y$.

39 Assume that the random variables X and Y have the joint distribution given in Table 4.6.

(a) What is $P(X \geq 1 \text{ and } Y \leq 0)$?

(b) What is the conditional probability that $Y \leq 0$ given that $X = 2$?

(c) Are X and Y independent?

(d) What is the distribution of $Z = XY$?

40 In the *problem of points*, discussed in the historical remarks in Section 3.2, two players, A and B, play a series of points in a game with player A winning each point with probability p and player B winning each point with probability $q = 1 - p$. The first player to win N points wins the game. Assume that $N = 3$. Let X be a random variable that has the value 1 if player A wins the series and 0 otherwise. Let Y be a random variable with value the number of points played in a game. Find the distribution of X and Y when $p = 1/2$. Are X and Y independent in this case? Answer the same questions for the case $p = 2/3$.

41 The letters between Pascal and Fermat, which are often credited with having started probability theory, dealt mostly with the *problem of points* described in Exercise 40. Pascal and Fermat considered the problem of finding a fair division of stakes if the game must be called off when the first player has won r games and the second player has won s games, with $r < N$ and $s < N$. Let $P(r, s)$ be the probability that player A wins the game if he has already won r points and player B has won s points. Then

(a) $P(r, N) = 0$ if $r < N$,

(b) $P(N, s) = 1$ if $s < N$,

(c) $P(r, s) = pP(r + 1, s) + qP(r, s + 1)$ if $r < N$ and $s < N$;

and (1), (2), and (3) determine $P(r, s)$ for $r \leq N$ and $s \leq N$. Pascal used these facts to find $P(r, s)$ by working backward: He first obtained $P(N-1, j)$ for $j = N-1, N-2, \ldots, 0$; then, from these values, he obtained $P(N-2, j)$ for $j = N-1, N-2, \ldots, 0$ and, continuing backward, obtained all the values $P(r, s)$. Write a program to compute $P(r, s)$ for given N, a, b, and p. *Warning*: Follow Pascal and you will be able to run $N = 100$; use recursion and you will not be able to run $N = 20$.

42 Fermat solved the *problem of points* (see Exercise 40) as follows: He realized that the problem was difficult because the possible ways the play might go are not equally likely. For example, when the first player needs two more games and the second needs three to win, two possible ways the series might go for the first player are WLW and LWLW. These sequences are not equally likely. To avoid this difficulty, Fermat extended the play, adding fictitious plays so that the series went the maximum number of games needed (four in this case). He obtained equally likely outcomes and used, in effect, the Pascal triangle to calculate $P(r, s)$. Show that this leads to a *formula* for $P(r, s)$ even for the case $p \neq 1/2$.

43 The Yankees are playing the Dodgers in a world series. The Yankees win each game with probability .6. What is the probability that the Yankees win the series? (The series is won by the first team to win four games.)

44 C. L. Anderson[11] has used Fermat's argument for the *problem of points* to prove the following result due to J. G. Kingston. You are playing the *game of points* (see Exercise 40) but, at each point, when you serve you win with probability p, and when your opponent serves you win with probability $\bar{p}$. You will serve first, but you can choose one of the following two conventions for serving: for the first convention you alternate service (tennis), and for the second the person serving continues to serve until he loses a point and then the other player serves (racquetball). The first player to win N points wins the game. The problem is to show that the probability of winning the game is the same under either convention.

(a) Show that, under either convention, you will serve at most N points and your opponent at most $N - 1$ points.

(b) Extend the number of points to $2N - 1$ so that you serve N points and your opponent serves $N - 1$. For example, you serve any additional points necessary to make N serves and then your opponent serves any additional points necessary to make him serve $N - 1$ points. The winner

[11] C. L. Anderson, "Note on the Advantage of First Serve," *Journal of Combinatorial Theory,* Series A, vol. 23 (1977), p. 363.

is now the person, in the extended game, who wins the most points. Show that playing these additional points has not changed the winner.

(c) Show that (a) and (b) prove that you have the same probability of winning the game under either convention.

45 In the previous problem, assume that $p = 1 - \bar{p}$.

(a) Show that under either service convention, the first player will win more often than the second player if and only if $p > .5$.

(b) In volleyball, a team can only win a point while it is serving. Thus, any individual "play" either ends with a point being awarded to the serving team or with the service changing to the other team. The first team to win N points wins the game. (We ignore here the additional restriction that the winning team must be ahead by at least two points at the end of the game.) Assume that each team has the same probability of winning the play when it is serving, i.e., that $p = 1 - \bar{p}$. Show that in this case, the team that serves first will win more than half the time, as long as $p > 0$. (If $p = 0$, then the game never ends.) *Hint*: Define p' to be the probability that a team wins the next point, given that it is serving. If we write $q = 1 - p$, then one can show that

$$p' = \frac{p}{1 - q^2} \ .$$

If one now considers this game in a slightly different way, one can see that the second service convention in the preceding problem can be used, with p replaced by p'.

46 A poker hand consists of 5 cards dealt from a deck of 52 cards. Let X and Y be, respectively, the number of aces and kings in a poker hand. Find the joint distribution of X and Y.

47 Let X_1 and X_2 be independent random variables and let $Y_1 = \phi_1(X_1)$ and $Y_2 = \phi_2(X_2)$.

(a) Show that

$$P(Y_1 = r, Y_2 = s) = \sum_{\substack{\phi_1(a)=r \\ \phi_2(b)=s}} P(X_1 = a, X_2 = b) \ .$$

(b) Using (a), show that $P(Y_1 = r, Y_2 = s) = P(Y_1 = r)P(Y_2 = s)$ so that Y_1 and Y_2 are independent.

48 Let Ω be the sample space of an experiment. Let E be an event with $P(E) > 0$ and define $m_E(\omega)$ by $m_E(\omega) = m(\omega|E)$. Prove that $m_E(\omega)$ is a distribution function on E, that is, that $m_E(\omega) \geq 0$ and that $\sum_{\omega \in \Omega} m_E(\omega) = 1$. The function m_E is called the *conditional distribution given* E.

49 You are given two urns each containing two biased coins. The coins in urn I come up heads with probability p_1, and the coins in urn II come up heads with probability $p_2 \neq p_1$. You are given a choice of (a) choosing an urn at random and tossing the two coins in this urn or (b) choosing one coin from each urn and tossing these two coins. You win a prize if both coins turn up heads. Show that you are better off selecting choice (a).

50 Prove that, if A_1, A_2, ..., A_n are independent events defined on a sample space Ω and if $0 < P(A_j) < 1$ for all j, then Ω must have at least 2^n points.

51 Prove that if

$$P(A|C) \geq P(B|C) \text{ and } P(A|\tilde{C}) \geq P(B|\tilde{C}) \ ,$$

then $P(A) \geq P(B)$.

52 A coin is in one of n boxes. The probability that it is in the ith box is p_i. If you search in the ith box and it is there, you find it with probability a_i. Show that the probability p that the coin is in the jth box, given that you have looked in the ith box and not found it, is

$$p = \begin{cases} p_j/(1 - a_i p_i), & \text{if } j \neq i, \\ (1 - a_i)p_i/(1 - a_i p_i), & \text{if } j = i. \end{cases}$$

53 George Wolford has suggested the following variation on the Linda problem (see Exercise 1.2.25). The registrar is carrying John and Mary's registration cards and drops them in a puddle. When he pickes them up he cannot read the names but on the first card he picked up he can make out Mathematics 23 and Government 35, and on the second card he can make out only Mathematics 23. He asks you if you can help him decide which card belongs to Mary. You know that Mary likes government but does not like mathematics. You know nothing about John and assume that he is just a typical Dartmouth student. From this you estimate:

$$\begin{aligned} P(\text{Mary takes Government 35}) &= .5 \ , \\ P(\text{Mary takes Mathematics 23}) &= .1 \ , \\ P(\text{John takes Government 35}) &= .3 \ , \\ P(\text{John takes Mathematics 23}) &= .2 \ . \end{aligned}$$

Assume that their choices for courses are independent events. Show that the card with Mathematics 23 and Government 35 showing is more likely to be Mary's than John's. The conjunction fallacy referred to in the Linda problem would be to assume that the event "Mary takes Mathematics 23 and Government 35" is more likely than the event "Mary takes Mathematics 23." Why are we not making this fallacy here?

54 (Suggested by Eisenberg and Ghosh[12]) A deck of playing cards can be described as a Cartesian product

$$\text{Deck} = \text{Suit} \times \text{Rank} ,$$

where Suit $= \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$ and Rank $= \{2, 3, \ldots, 10, \text{J}, \text{Q}, \text{K}, \text{A}\}$. This just means that every card may be thought of as an ordered pair like $(\diamondsuit, 2)$. By a *suit event* we mean any event A contained in Deck which is described in terms of Suit alone. For instance, if A is "the suit is red," then

$$A = \{\diamondsuit, \heartsuit\} \times \text{Rank} ,$$

so that A consists of all cards of the form $(\diamondsuit, r)$ or $(\heartsuit, r)$ where r is any rank. Similarly, a *rank event* is any event described in terms of rank alone.

(a) Show that if A is any suit event and B any rank event, then A and B are *independent.* (We can express this briefly by saying that suit and rank are independent.)

(b) Throw away the ace of spades. Show that now no nontrivial (i.e., neither empty nor the whole space) suit event A is independent of any nontrivial rank event B. *Hint*: Here independence comes down to

$$c/51 = (a/51) \cdot (b/51) ,$$

where a, b, c are the respective sizes of A, B and $A \cap B$. It follows that 51 must divide ab, hence that 3 must divide one of a and b, and 17 the other. But the possible sizes for suit and rank events preclude this.

(c) Show that the deck in (b) nevertheless does have pairs A, B of nontrivial independent events. *Hint*: Find 2 events A and B of sizes 3 and 17, respectively, which intersect in a single point.

(d) Add a joker to a full deck. Show that now there is no pair A, B of nontrivial independent events. *Hint*: See the hint in (b); 53 is prime.

The following problems are suggested by Stanley Gudder in his article "Do Good Hands Attract?"[13] He says that event A *attracts* event B if $P(B|A) > P(B)$ and *repels* B if $P(B|A) < P(B)$.

55 Let R_i be the event that the ith player in a poker game has a royal flush. Show that a royal flush (A,K,Q,J,10 of one suit) attracts another royal flush, that is $P(R_2|R_1) > P(R_2)$. Show that a royal flush repels full houses.

56 Prove that A attracts B if and only if B attracts A. Hence we can say that A and B are *mutually attractive* if A attracts B.

[12] B. Eisenberg and B. K. Ghosh, "Independent Events in a Discrete Uniform Probability Space," *The American Statistician,* vol. 41, no. 1 (1987), pp. 52–56.

[13] S. Gudder, "Do Good Hands Attract?" *Mathematics Magazine,* vol. 54, no. 1 (1981), pp. 13–16.

57 Prove that A neither attracts nor repels B if and only if A and B are independent.

58 Prove that A and B are mutually attractive if and only if $P(B|A) > P(B|\tilde{A})$.

59 Prove that if A attracts B, then A repels $\tilde{B}$.

60 Prove that if A attracts both B and C, and A repels $B \cap C$, then A attracts $B \cup C$. Is there any example in which A attracts both B and C and repels $B \cup C$?

61 Prove that if $B_1, B_2, \ldots, B_n$ are mutually disjoint and collectively exhaustive, and if A attracts some B_i, then A must repel some B_j.

62 (a) Suppose that you are looking in your desk for a letter from some time ago. Your desk has eight drawers, and you assess the probability that it is in any particular drawer is 10% (so there is a 20% chance that it is not in the desk at all). Suppose now that you start searching systematically through your desk, one drawer at a time. In addition, suppose that you have not found the letter in the first i drawers, where $0 \leq i \leq 7$. Let p_i denote the probability that the letter will be found in the next drawer, and let q_i denote the probability that the letter will be found in some subsequent drawer (both p_i and q_i are conditional probabilities, since they are based upon the assumption that the letter is not in the first i drawers). Show that the p_i's increase and the q_i's decrease. (This problem is from Falk et al.[14])

(b) The following data appeared in an article in the Wall Street Journal.[15] For the ages 20, 30, 40, 50, and 60, the probability of a woman in the U.S. developing cancer in the next ten years is 0.5%, 1.2%, 3.2%, 6.4%, and 10.8%, respectively. At the same set of ages, the probability of a woman in the U.S. eventually developing cancer is 39.6%, 39.5%, 39.1%, 37.5%, and 34.2%, respectively. Do you think that the problem in part (a) gives an explanation for these data?

63 Here are two variations of the Monty Hall problem that are discussed by Granberg.[16]

(a) Suppose that everything is the same except that Monty forgot to find out in advance which door has the car behind it. In the spirit of "the show must go on," he makes a guess at which of the two doors to open and gets lucky, opening a door behind which stands a goat. Now should the contestant switch?

[14]R. Falk, A. Lipson, and C. Konold, "The ups and downs of the hope function in a fruitless search," in *Subjective Probability,* G. Wright and P. Ayton, (eds.) (Chichester: Wiley, 1994), pgs. 353-377.

[15]C. Crossen, "Fright by the numbers: Alarming disease data are frequently flawed," *Wall Street Journal,* 11 April 1996, p. B1.

[16]D. Granberg, "To switch or not to switch," in *The power of logical thinking,* M. vos Savant, (New York: St. Martin's 1996).

(b) You have observed the show for a long time and found that the car is put behind door A 45% of the time, behind door B 40% of the time and behind door C 15% of the time. Assume that everything else about the show is the same. Again you pick door A. Monty opens a door with a goat and offers to let you switch. Should you? Suppose you knew in advance that Monty was going to give you a chance to switch. Should you have initially chosen door A?

4.2 Continuous Conditional Probability

In situations where the sample space is continuous we will follow the same procedure as in the previous section. Thus, for example, if X is a continuous random variable with density function $f(x)$, and if E is an event with positive probability, we define a conditional density function by the formula

$$f(x|E) = \begin{cases} f(x)/P(E), & \text{if } x \in E, \\ 0, & \text{if } x \notin E. \end{cases}$$

Then for any event F, we have

$$P(F|E) = \int_F f(x|E)\,dx \ .$$

The expression $P(F|E)$ is called the conditional probability of F given E. As in the previous section, it is easy to obtain an alternative expression for this probability:

$$P(F|E) = \int_F f(x|E)\,dx = \int_{E\cap F} \frac{f(x)}{P(E)}\,dx = \frac{P(E\cap F)}{P(E)} \ .$$

We can think of the conditional density function as being 0 except on E, and normalized to have integral 1 over E. Note that if the original density is a uniform density corresponding to an experiment in which all events of equal size are *equally likely,* then the same will be true for the conditional density.

Example 4.18 In the spinner experiment (cf. Example 2.1), suppose we know that the spinner has stopped with head in the upper half of the circle, $0 \le x \le 1/2$. What is the probability that $1/6 \le x \le 1/3$?

Here $E = [0, 1/2]$, $F = [1/6, 1/3]$, and $F \cap E = F$. Hence

$$\begin{aligned} P(F|E) &= \frac{P(F\cap E)}{P(E)} \\ &= \frac{1/6}{1/2} \\ &= \frac{1}{3} \ , \end{aligned}$$

which is reasonable, since F is $1/3$ the size of E. The conditional density function here is given by

$$f(x|E) = \begin{cases} 2, & \text{if } 0 \le x < 1/2, \\ 0, & \text{if } 1/2 \le x < 1. \end{cases}$$

Thus the conditional density function is nonzero only on $[0, 1/2]$, and is uniform there. □

Example 4.19 In the dart game (cf. Example 2.8), suppose we know that the dart lands in the upper half of the target. What is the probability that its distance from the center is less than 1/2?

Here $E = \{\, (x, y) : y \ge 0 \,\}$, and $F = \{\, (x, y) : x^2 + y^2 < (1/2)^2 \,\}$. Hence,

$$\begin{aligned} P(F|E) &= \frac{P(F \cap E)}{P(E)} = \frac{(1/\pi)[(1/2)(\pi/4)]}{(1/\pi)(\pi/2)} \\ &= 1/4 \ . \end{aligned}$$

Here again, the size of $F \cap E$ is 1/4 the size of E. The conditional density function is

$$f((x, y)|E) = \begin{cases} f(x, y)/P(E) = 2/\pi, & \text{if } (x, y) \in E, \\ 0, & \text{if } (x, y) \notin E. \end{cases}$$

□

Example 4.20 We return to the exponential density (cf. Example 2.17). We suppose that we are observing a lump of plutonium-239. Our experiment consists of waiting for an emission, then starting a clock, and recording the length of time X that passes until the next emission. Experience has shown that X has an exponential density with some parameter λ, which depends upon the size of the lump. Suppose that when we perform this experiment, we notice that the clock reads r seconds, and is still running. What is the probability that there is no emission in a further s seconds?

Let $G(t)$ be the probability that the next particle is emitted after time t. Then

$$\begin{aligned} G(t) &= \int_t^\infty \lambda e^{-\lambda x}\, dx \\ &= -e^{-\lambda x}\Big|_t^\infty = e^{-\lambda t} \ . \end{aligned}$$

Let E be the event "the next particle is emitted after time r" and F the event "the next particle is emitted after time $r + s$." Then

$$\begin{aligned} P(F|E) &= \frac{P(F \cap E)}{P(E)} \\ &= \frac{G(r+s)}{G(r)} \\ &= \frac{e^{-\lambda(r+s)}}{e^{-\lambda r}} \\ &= e^{-\lambda s} \ . \end{aligned}$$

This tells us the rather surprising fact that the probability that we have to wait s seconds more for an emission, given that there has been no emission in r seconds, is *independent* of the time r. This property (called the *memoryless* property) was introduced in Example 2.17. When trying to model various phenomena, this property is helpful in deciding whether the exponential density is appropriate.

The fact that the exponential density is memoryless means that it is reasonable to assume if one comes upon a lump of a radioactive isotope at some random time, then the amount of time until the next emission has an exponential density with the same parameter as the time between emissions. A well-known example, known as the "bus paradox," replaces the emissions by buses. The apparent paradox arises from the following two facts: 1) If you know that, on the average, the buses come by every 30 minutes, then if you come to the bus stop at a random time, you should only have to wait, on the average, for 15 minutes for a bus, and 2) Since the buses arrival times are being modelled by the exponential density, then no matter when you arrive, you will have to wait, on the average, for 30 minutes for a bus.

The reader can now see that in Exercises 2.2.9, 2.2.10, and 2.2.11, we were asking for simulations of conditional probabilities, under various assumptions on the distribution of the interarrival times. If one makes a reasonable assumption about this distribution, such as the one in Exercise 2.2.10, then the average waiting time is more nearly one-half the average interarrival time. □

Independent Events

If E and F are two events with positive probability in a continuous sample space, then, as in the case of discrete sample spaces, we define E and F to be *independent* if $P(E|F) = P(E)$ and $P(F|E) = P(F)$. As before, each of the above equations imply the other, so that to see whether two events are independent, only one of these equations must be checked. It is also the case that, if E and F are independent, then $P(E \cap F) = P(E)P(F)$.

Example 4.21 (Example 4.18 continued) In the dart game (see Example 4.18), let E be the event that the dart lands in the *upper* half of the target ($y \geq 0$) and F the event that the dart lands in the *right* half of the target ($x \geq 0$). Then $P(E \cap F)$ is the probability that the dart lies in the first quadrant of the target, and

$$\begin{aligned} P(E \cap F) &= \frac{1}{\pi} \int_{E \cap F} 1 \, dxdy \\ &= \text{Area}\,(E \cap F) \\ &= \text{Area}\,(E)\,\text{Area}\,(F) \\ &= \left(\frac{1}{\pi} \int_E 1 \, dxdy\right) \left(\frac{1}{\pi} \int_F 1 \, dxdy\right) \\ &= P(E)P(F) \end{aligned}$$

so that E and F are independent. What makes this work is that the events E and F are described by restricting different coordinates. This idea is made more precise below. □

Joint Density and Cumulative Distribution Functions

In a manner analogous with discrete random variables, we can define joint density functions and cumulative distribution functions for multi-dimensional continuous random variables.

Definition 4.6 Let $X_1,\ X_2, \ldots,\ X_n$ be continuous random variables associated with an experiment, and let $\bar{X} = (X_1,\ X_2, \ldots,\ X_n)$. Then the joint cumulative distribution function of $\bar{X}$ is defined by

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \ .$$

The joint density function of $\bar{X}$ satisfies the following equation:

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(t_1, t_2, \ldots t_n)\, dt_n dt_{n-1} \ldots dt_1 \ .$$

□

It is straightforward to show that, in the above notation,

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n} \ . \tag{4.4}$$

Independent Random Variables

As with discrete random variables, we can define mutual independence of continuous random variables.

Definition 4.7 Let X_1, X_2, ..., X_n be continuous random variables with cumulative distribution functions $F_1(x),\ F_2(x), \ldots,\ F_n(x)$. Then these random variables are *mutually independent* if

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1) F_2(x_2) \cdots F_n(x_n)$$

for any choice of $x_1, x_2, \ldots, x_n$. Thus, if $X_1,\ X_2, \ldots,\ X_n$ are mutually independent, then the joint cumulative distribution function of the random variable $\bar{X} = (X_1, X_2, \ldots, X_n)$ is just the product of the individual cumulative distribution functions. When two random variables are mutually independent, we shall say more briefly that they are *independent.* □

Using Equation 4.4, the following theorem can easily be shown to hold for mutually independent continuous random variables.

Theorem 4.2 Let X_1, X_2, ..., X_n be continuous random variables with density functions $f_1(x),\ f_2(x), \ldots,\ f_n(x)$. Then these random variables are *mutually independent* if and only if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$$

for any choice of $x_1, x_2, \ldots, x_n$. □

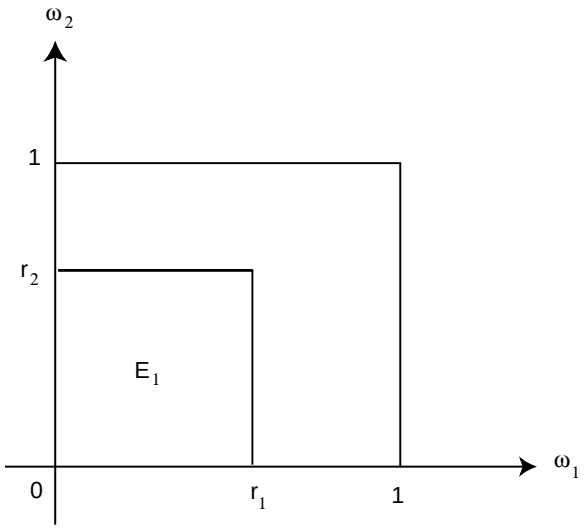

Figure 4.7: X_1 and X_2 are independent.

Let's look at some examples.

Example 4.22 In this example, we define three random variables, X_1, X_2, and X_3. We will show that X_1 and X_2 are independent, and that X_1 and X_3 are not independent. Choose a point $\omega = (\omega_1, \omega_2)$ at random from the unit square. Set $X_1 = \omega_1^2$, $X_2 = \omega_2^2$, and $X_3 = \omega_1 + \omega_2$. Find the joint distributions $F_{12}(r_1, r_2)$ and $F_{23}(r_2, r_3)$.

We have already seen (see Example 2.13) that

$$\begin{aligned} F_1(r_1) &= P(-\infty < X_1 \leq r_1) \\ &= \sqrt{r_1}, \qquad \text{if } 0 \leq r_1 \leq 1 \ , \end{aligned}$$

and similarly,

$$F_2(r_2) = \sqrt{r_2} \ ,$$

if $0 \leq r_2 \leq 1$. Now we have (see Figure 4.7)

$$\begin{aligned} F_{12}(r_1, r_2) &= P(X_1 \leq r_1 \text{ and } X_2 \leq r_2) \\ &= P(\omega_1 \leq \sqrt{r_1} \text{ and } \omega_2 \leq \sqrt{r_2}) \\ &= \text{Area}\,(E_1) \\ &= \sqrt{r_1}\sqrt{r_2} \\ &= F_1(r_1)F_2(r_2) \ . \end{aligned}$$

In this case $F_{12}(r_1, r_2) = F_1(r_1)F_2(r_2)$ so that X_1 and X_2 are independent. On the other hand, if $r_1 = 1/4$ and $r_3 = 1$, then (see Figure 4.8)

$$F_{13}(1/4, 1) = P(X_1 \leq 1/4, \ X_3 \leq 1)$$

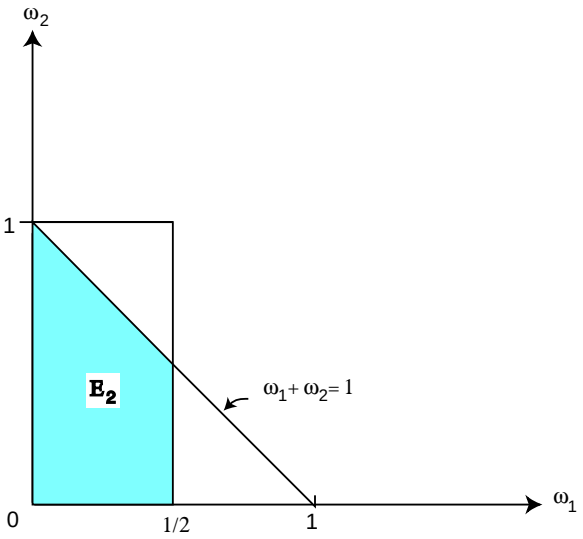

Figure 4.8: X_1 and X_3 are not independent.

$$\begin{aligned} &= P(\omega_1 \leq 1/2,\ \omega_1 + \omega_2 \leq 1) \\ &= \text{Area}\,(E_2) \\ &= \frac{1}{2} - \frac{1}{8} = \frac{3}{8}\,. \end{aligned}$$

Now recalling that

$$F_3(r_3) = \begin{cases} 0, & \text{if } r_3 < 0, \\ (1/2)r_3^2, & \text{if } 0 \leq r_3 \leq 1, \\ 1 - (1/2)(2 - r_3)^2, & \text{if } 1 \leq r_3 \leq 2, \\ 1, & \text{if } 2 < r_3, \end{cases}$$

(see Example 2.14), we have $F_1(1/4)F_3(1) = (1/2)(1/2) = 1/4$. Hence, X_1 and X_3 are not independent random variables. A similar calculation shows that X_2 and X_3 are not independent either. □

Although we shall not prove it here, the following theorem is a useful one. The statement also holds for mutually independent discrete random variables. A proof may be found in Rényi.[17]

Theorem 4.3 Let $X_1, X_2, \ldots, X_n$ be mutually independent continuous random variables and let $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$ be continuous functions. Then $\phi_1(X_1)$, $\phi_2(X_2), \ldots, \phi_n(X_n)$ are mutually independent. □

Independent Trials

Using the notion of independence, we can now formulate for continuous sample spaces the notion of independent trials (see Definition 4.5).

[17] A. Rényi, *Probability Theory* (Budapest: Akadémiai Kiadó, 1970), p. 183.

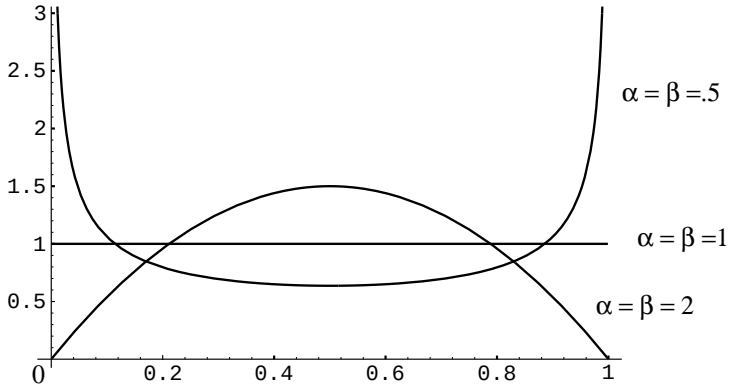

Figure 4.9: Beta density for $\alpha = \beta = .5, 1, 2$.

Definition 4.8 A sequence X_1, X_2, ..., X_n of random variables X_i that are mutually independent and have the same density is called an *independent trials process.* □

As in the case of discrete random variables, these independent trials processes arise naturally in situations where an experiment described by a single random variable is repeated n times.

Beta Density

We consider next an example which involves a sample space with both discrete and continuous coordinates. For this example we shall need a new density function called the *beta density.* This density has two parameters α, β and is defined by

$$B(\alpha, \beta, x) = \begin{cases} (1/B(\alpha, \beta))x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Here α and β are any positive numbers, and the beta function $B(\alpha, \beta)$ is given by the area under the graph of $x^{\alpha-1}(1-x)^{\beta-1}$ between 0 and 1:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx \ .$$

Note that when $\alpha = \beta = 1$ the beta density if the uniform density. When α and β are greater than 1 the density is bell-shaped, but when they are less than 1 it is U-shaped as suggested by the examples in Figure 4.9.

We shall need the values of the beta function only for integer values of α and β, and in this case

$$B(\alpha, \beta) = \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \ .$$

Example 4.23 In medical problems it is often assumed that a drug is effective with a probability x each time it is used and the various trials are independent, so that

one is, in effect, tossing a biased coin with probability x for heads. Before further experimentation, you do not know the value x but past experience might give some information about its possible values. It is natural to represent this information by sketching a density function to determine a distribution for x. Thus, we are considering x to be a continuous random variable, which takes on values between 0 and 1. If you have no knowledge at all, you would sketch the uniform density. If past experience suggests that x is very likely to be near 2/3 you would sketch a density with maximum at 2/3 and a spread reflecting your uncertainly in the estimate of 2/3. You would then want to find a density function that reasonably fits your sketch. The beta densities provide a class of densities that can be fit to most sketches you might make. For example, for $\alpha > 1$ and $\beta > 1$ it is bell-shaped with the parameters α and β determining its peak and its spread.

Assume that the experimenter has chosen a beta density to describe the state of his knowledge about x before the experiment. Then he gives the drug to n subjects and records the number i of successes. The number i is a discrete random variable, so we may conveniently describe the set of possible outcomes of this experiment by referring to the ordered pair (x, i).

We let $m(i|x)$ denote the probability that we observe i successes given the value of x. By our assumptions, $m(i|x)$ is the binomial distribution with probability x for success:

$$m(i|x) = b(n, x, i) = \binom{n}{i} x^i (1-x)^j \ ,$$

where $j = n - i$.

If x is chosen at random from $[0, 1]$ with a beta density $B(\alpha, \beta, x)$, then the density function for the outcome of the pair (x, i) is

$$\begin{aligned} f(x,i) &= m(i|x)B(\alpha,\beta,x) \\ &= \binom{n}{i} x^i (1-x)^j \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ &= \binom{n}{i} \frac{1}{B(\alpha,\beta)} x^{\alpha+i-1}(1-x)^{\beta+j-1} \ . \end{aligned}$$

Now let $m(i)$ be the probability that we observe i successes *not* knowing the value of x. Then

$$\begin{aligned} m(i) &= \int_0^1 m(i|x)B(\alpha,\beta,x)\,dx \\ &= \binom{n}{i} \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+i-1}(1-x)^{\beta+j-1}\,dx \\ &= \binom{n}{i} \frac{B(\alpha+i,\beta+j)}{B(\alpha,\beta)} \ . \end{aligned}$$

Hence, the probability density $f(x|i)$ for x, given that i successes were observed, is

$$f(x|i) = \frac{f(x,i)}{m(i)}$$

$$= \frac{x^{\alpha+i-1}(1-x)^{\beta+j-1}}{B(\alpha+i,\beta+j)} \ , \tag{4.5}$$

that is, $f(x|i)$ is another beta density. This says that if we observe i successes and j failures in n subjects, then the new density for the probability that the drug is effective is again a beta density but with parameters $\alpha+i$, $\beta+j$.

Now we assume that before the experiment we choose a beta density with parameters α and β, and that in the experiment we obtain i successes in n trials. We have just seen that in this case, the new density for x is a beta density with parameters $\alpha+i$ and $\beta+j$.

Now we wish to calculate the probability that the drug is effective on the next subject. For any particular real number t between 0 and 1, the probability that x has the value t is given by the expression in Equation 4.5. Given that x has the value t, the probability that the drug is effective on the next subject is just t. Thus, to obtain the probability that the drug is effective on the next subject, we integrate the product of the expression in Equation 4.5 and t over all possible values of t. We obtain:

$$\begin{aligned}
&\frac{1}{B(\alpha+i,\beta+j)} \int_0^1 t \cdot t^{\alpha+i-1}(1-t)^{\beta+j-1}\,dt \\
&= \frac{B(\alpha+i+1,\beta+j)}{B(\alpha+i,\beta+j)} \\
&= \frac{(\alpha+i)!\,(\beta+j-1)!}{(\alpha+\beta+i+j)!} \cdot \frac{(\alpha+\beta+i+j-1)!}{(\alpha+i-1)!\,(\beta+j-1)!} \\
&= \frac{\alpha+i}{\alpha+\beta+n} \ .
\end{aligned}$$

If n is large, then our estimate for the probability of success after the experiment is approximately the proportion of successes observed in the experiment, which is certainly a reasonable conclusion. □

The next example is another in which the true probabilities are unknown and must be estimated based upon experimental data.

Example 4.24 (Two-armed bandit problem) You are in a casino and confronted by two slot machines. Each machine pays off either 1 dollar or nothing. The probability that the first machine pays off a dollar is x and that the second machine pays off a dollar is y. We assume that x and y are random numbers chosen independently from the interval $[0,1]$ and unknown to you. You are permitted to make a series of ten plays, each time choosing one machine or the other. How should you choose to maximize the number of times that you win?

One strategy that sounds reasonable is to calculate, at every stage, the probability that each machine will pay off and choose the machine with the higher probability. Let win(i), for $i = 1$ or 2, be the number of times that you have won on the ith machine. Similarly, let lose(i) be the number of times you have lost on the ith machine. Then, from Example 4.23, the probability $p(i)$ that you win if you

Machine	Result
1	W
1	L
2	L
1	L
1	W
1	L
1	L
1	L
2	W
2	L

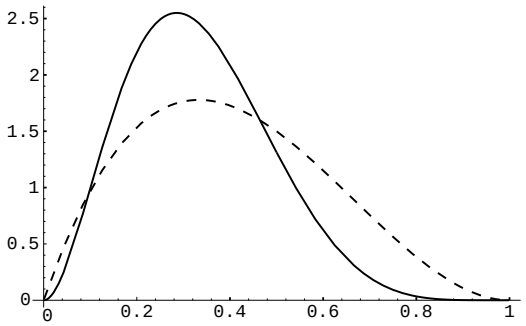

Figure 4.10: Play the best machine.

choose the ith machine is

$$p(i) = \frac{\text{win}(i) + 1}{\text{win}(i) + \text{lose}(i) + 2} \ .$$

Thus, if $p(1) > p(2)$ you would play machine 1 and otherwise you would play machine 2. We have written a program **TwoArm** to simulate this experiment. In the program, the user specifies the initial values for x and y (but these are unknown to the experimenter). The program calculates at each stage the two conditional densities for x and y, given the outcomes of the previous trials, and then computes $p(i)$, for $i = 1, 2$. It then chooses the machine with the highest value for the probability of winning for the next play. The program prints the machine chosen on each play and the outcome of this play. It also plots the new densities for x (solid line) and y (dotted line), showing only the current densities. We have run the program for ten plays for the case $x = .6$ and $y = .7$. The result is shown in Figure 4.10.

The run of the program shows the weakness of this strategy. Our initial probability for winning on the better of the two machines is .7. We start with the poorer machine and our outcomes are such that we always have a probability greater than .6 of winning and so we just keep playing this machine even though the other machine is better. If we had lost on the first play we would have switched machines. Our final density for y is the same as our initial density, namely, the uniform density. Our final density for x is different and reflects a much more accurate knowledge about x. The computer did pretty well with this strategy, winning seven out of the ten trials, but ten trials are not enough to judge whether this is a good strategy in the long run.

Another popular strategy is the *play-the-winner strategy.* As the name suggests, for this strategy we choose the same machine when we win and switch machines when we lose. The program **TwoArm** will simulate this strategy as well. In Figure 4.11, we show the results of running this program with the play-the-winner strategy and the same true probabilities of .6 and .7 for the two machines. After ten plays our densities for the unknown probabilities of winning suggest to us that the second machine is indeed the better of the two. We again won seven out of the ten trials.

Machine	Result
1	W
1	W
1	L
2	L
1	W
1	W
1	L
2	L
1	L
2	W

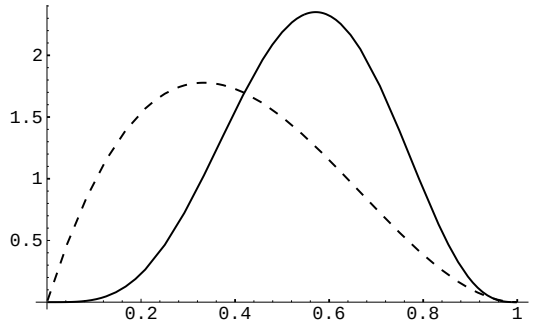

Figure 4.11: Play the winner.

Neither of the strategies that we simulated is the best one in terms of maximizing our average winnings. This best strategy is very complicated but is reasonably approximated by the play-the-winner strategy. Variations on this example have played an important role in the problem of clinical tests of drugs where experimenters face a similar situation. □

Exercises

1 Pick a point x at random (with uniform density) in the interval $[0, 1]$. Find the probability that $x > 1/2$, given that

(a) $x > 1/4$.

(b) $x < 3/4$.

(c) $|x - 1/2| < 1/4$.

(d) $x^2 - x + 2/9 < 0$.

2 A radioactive material emits α-particles at a rate described by the density function

$$f(t) = .1e^{-.1t} .$$

Find the probability that a particle is emitted in the first 10 seconds, given that

(a) no particle is emitted in the first second.

(b) no particle is emitted in the first 5 seconds.

(c) a particle is emitted in the first 3 seconds.

(d) a particle is emitted in the first 20 seconds.

3 The Acme Super light bulb is known to have a useful life described by the density function

$$f(t) = .01e^{-.01t} ,$$

where time t is measured in hours.

(a) Find the *failure rate* of this bulb (see Exercise 2.2.6).

(b) Find the *reliability* of this bulb after 20 hours.

(c) Given that it lasts 20 hours, find the probability that the bulb lasts another 20 hours.

(d) Find the probability that the bulb burns out in the forty-first hour, given that it lasts 40 hours.

4 Suppose you toss a dart at a circular target of radius 10 inches. Given that the dart lands in the upper half of the target, find the probability that

(a) it lands in the right half of the target.

(b) its distance from the center is less than 5 inches.

(c) its distance from the center is greater than 5 inches.

(d) it lands within 5 inches of the point $(0, 5)$.

5 Suppose you choose two numbers x and y, independently at random from the interval $[0, 1]$. Given that their sum lies in the interval $[0, 1]$, find the probability that

(a) $|x - y| < 1$.

(b) $xy < 1/2$.

(c) $\max\{x, y\} < 1/2$.

(d) $x^2 + y^2 < 1/4$.

(e) $x > y$.

6 Find the conditional density functions for the following experiments.

(a) A number x is chosen at random in the interval $[0, 1]$, given that $x > 1/4$.

(b) A number t is chosen at random in the interval $[0, \infty)$ with exponential density e^{-t}, given that $1 < t < 10$.

(c) A dart is thrown at a circular target of radius 10 inches, given that it falls in the upper half of the target.

(d) Two numbers x and y are chosen at random in the interval $[0, 1]$, given that $x > y$.

7 Let x and y be chosen at random from the interval $[0, 1]$. Show that the events $x > 1/3$ and $y > 2/3$ are independent events.

8 Let x and y be chosen at random from the interval $[0, 1]$. Which pairs of the following events are independent?

(a) $x > 1/3$.

(b) $y > 2/3$.

(c) $x > y$.

(d) $x + y < 1$.

9 Suppose that X and Y are continuous random variables with density functions $f_X(x)$ and $f_Y(y)$, respectively. Let $f(x, y)$ denote the joint density function of (X, Y). Show that

$$\int_{-\infty}^{\infty} f(x, y)\, dy = f_X(x) \ ,$$

and

$$\int_{-\infty}^{\infty} f(x, y)\, dx = f_Y(y) \ .$$

***10** In Exercise 2.2.12 you proved the following: If you take a stick of unit length and break it into three pieces, choosing the breaks at random (i.e., choosing two real numbers independently and uniformly from [0, 1]), then the probability that the three pieces form a triangle is 1/4. Consider now a similar experiment: First break the stick at random, then break the longer piece at random. Show that the two experiments are actually quite different, as follows:

(a) Write a program which simulates both cases for a run of 1000 trials, prints out the proportion of successes for each run, and repeats this process ten times. (Call a trial a success if the three pieces do form a triangle.) Have your program pick (x, y) at random in the unit square, and in each case use x and y to find the two breaks. For each experiment, have it plot (x, y) if (x, y) gives a success.

(b) Show that in the second experiment the theoretical probability of success is actually $2 \log 2 - 1$.

11 A coin has an unknown bias p that is assumed to be uniformly distributed between 0 and 1. The coin is tossed n times and heads turns up j times and tails turns up k times. We have seen that the probability that heads turns up next time is

$$\frac{j+1}{n+2} \ .$$

Show that this is the same as the probability that the next ball is black for the Polya urn model of Exercise 4.1.20. Use this result to explain why, in the Polya urn model, the proportion of black balls does not tend to 0 or 1 as one might expect but rather to a uniform distribution on the interval $[0, 1]$.

12 Previous experience with a drug suggests that the probability p that the drug is effective is a random quantity having a beta density with parameters $\alpha = 2$ and $\beta = 3$. The drug is used on ten subjects and found to be successful in four out of the ten patients. What density should we now assign to the probability p? What is the probability that the drug will be successful the next time it is used?

13 Write a program to allow you to compare the strategies play-the-winner and play-the-best-machine for the two-armed bandit problem of Example 4.24. Have your program determine the initial payoff probabilities for each machine by choosing a pair of random numbers between 0 and 1. Have your program carry out 20 plays and keep track of the number of wins for each of the two strategies. Finally, have your program make 1000 repetitions of the 20 plays and compute the average winning per 20 plays. Which strategy seems to be the best? Repeat these simulations with 20 replaced by 100. Does your answer to the above question change?

14 Consider the two-armed bandit problem of Example 4.24. Bruce Barnes proposed the following strategy, which is a variation on the play-the-best-machine strategy. The machine with the greatest probability of winning is played *unless* the following two conditions hold: (a) the difference in the probabilities for winning is less than .08, and (b) the ratio of the number of times played on the more often played machine to the number of times played on the less often played machine is greater than 1.4. If the above two conditions hold, then the machine with the smaller probability of winning is played. Write a program to simulate this strategy. Have your program choose the initial payoff probabilities at random from the unit interval $[0, 1]$, make 20 plays, and keep track of the number of wins. Repeat this experiment 1000 times and obtain the average number of wins per 20 plays. Implement a second strategy—for example, play-the-best-machine or one of your own choice, and see how this second strategy compares with Bruce's on average wins.

4.3 Paradoxes

Much of this section is based on an article by Snell and Vanderbei.[18]

One must be very careful in dealing with problems involving conditional probability. The reader will recall that in the Monty Hall problem (Example 4.6), if the contestant chooses the door with the car behind it, then Monty has a choice of doors to open. We made an assumption that in this case, he will choose each door with probability 1/2. We then noted that if this assumption is changed, the answer to the original question changes. In this section, we will study other examples of the same phenomenon.

Example 4.25 Consider a family with two children. Given that one of the children is a boy, what is the probability that both children are boys?

One way to approach this problem is to say that the other child is equally likely to be a boy or a girl, so the probability that both children are boys is 1/2. The "textbook" solution would be to draw the tree diagram and then form the conditional tree by deleting paths to leave only those paths that are consistent with the given

[18] J. L. Snell and R. Vanderbei, "Three Bewitching Paradoxes," in *Topics in Contemporary Probability and Its Applications*, CRC Press, Boca Raton, 1995.

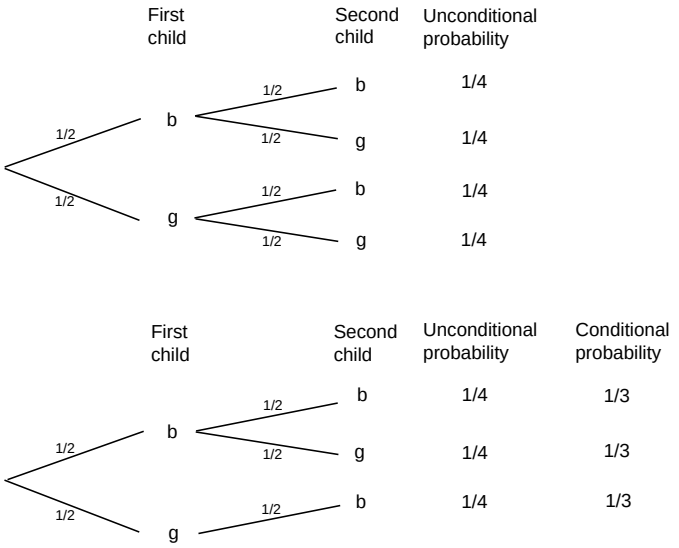

Figure 4.12: Tree for Example 4.25.

information. The result is shown in Figure 4.12. We see that the probability of two boys given a boy in the family is not 1/2 but rather 1/3. □

This problem and others like it are discussed in Bar-Hillel and Falk.[19] These authors stress that the answer to conditional probabilities of this kind can change depending upon how the information given was actually obtained. For example, they show that 1/2 is the correct answer for the following scenario.

Example 4.26 Mr. Smith is the father of two. We meet him walking along the street with a young boy whom he proudly introduces as his son. What is the probability that Mr. Smith's other child is also a boy?

As usual we have to make some additional assumptions. For example, we will assume that if Mr. Smith has a boy and a girl, he is equally likely to choose either one to accompany him on his walk. In Figure 4.13 we show the tree analysis of this problem and we see that 1/2 is, indeed, the correct answer. □

Example 4.27 It is not so easy to think of reasonable scenarios that would lead to the classical 1/3 answer. An attempt was made by Stephen Geller in proposing this problem to Marilyn vos Savant.[20] Geller's problem is as follows: A shopkeeper says she has two new baby beagles to show you, but she doesn't know whether they're both male, both female, or one of each sex. You tell her that you want only a male, and she telephones the fellow who's giving them a bath. "Is at least one a male?"

[19]M. Bar-Hillel and R. Falk, "Some teasers concerning conditional probabilities," *Cognition*, vol. 11 (1982), pgs. 109-122.

[20]M. vos Savant, "Ask Marilyn," *Parade Magazine*, 9 September; 2 December; 17 February 1990, reprinted in Marilyn vos Savant, *Ask Marilyn*, St. Martins, New York, 1992.

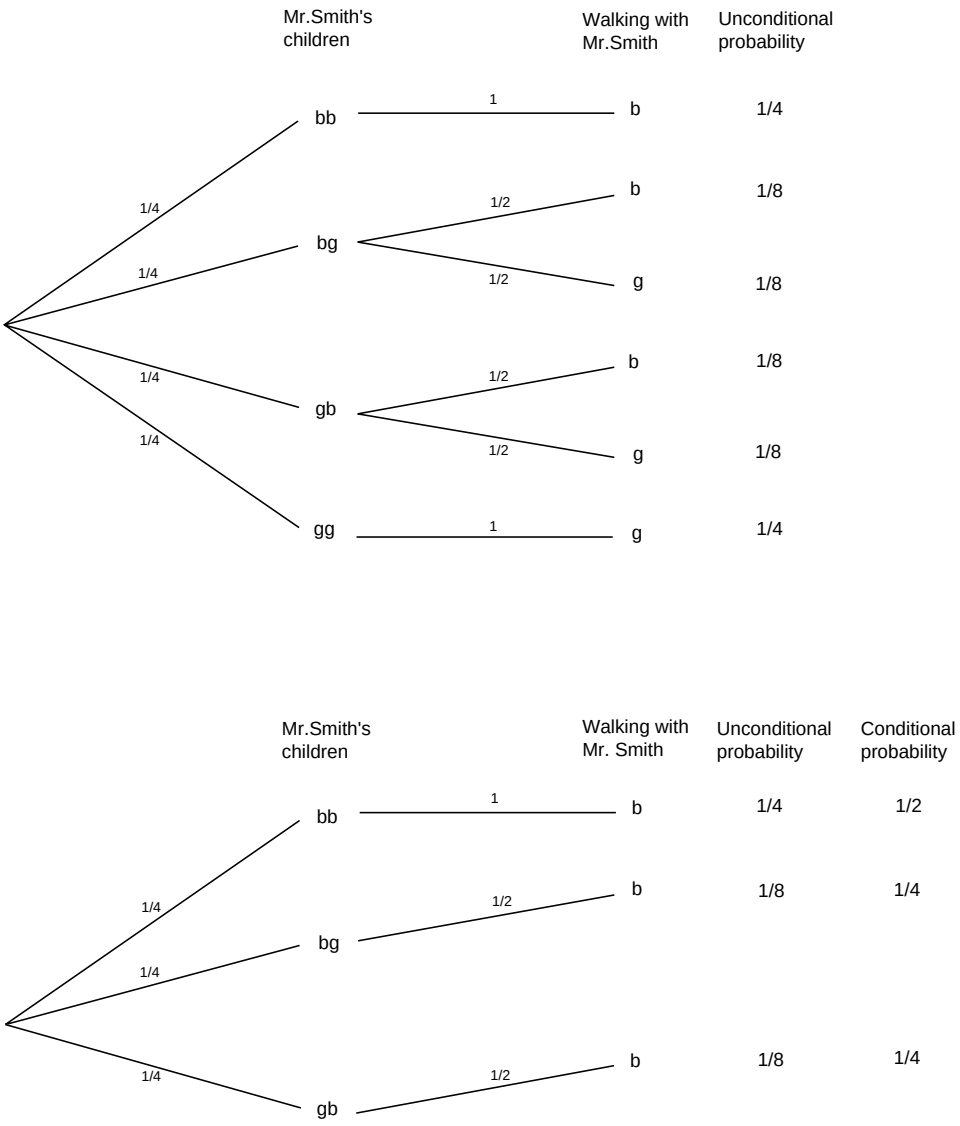

Figure 4.13: Tree for Example 4.26.

she asks. "Yes," she informs you with a smile. What is the probability that the *other* one is male?

The reader is asked to decide whether the model which gives an answer of 1/3 is a reasonable one to use in this case. □

In the preceding examples, the apparent paradoxes could easily be resolved by clearly stating the model that is being used and the assumptions that are being made. We now turn to some examples in which the paradoxes are not so easily resolved.

Example 4.28 Two envelopes each contain a certain amount of money. One envelope is given to Ali and the other to Baba and they are told that one envelope contains twice as much money as the other. However, neither knows who has the larger prize. Before anyone has opened their envelope, Ali is asked if she would like to trade her envelope with Baba. She reasons as follows: Assume that the amount in my envelope is x. If I switch, I will end up with $x/2$ with probability 1/2, and $2x$ with probability 1/2. If I were given the opportunity to play this game many times, and if I were to switch each time, I would, on average, get

$$\frac{1}{2}\frac{x}{2} + \frac{1}{2}2x = \frac{5}{4}x \ .$$

This is greater than my average winnings if I didn't switch.

Of course, Baba is presented with the same opportunity and reasons in the same way to conclude that he too would like to switch. So they switch and each thinks that his/her net worth just went up by 25%.

Since neither has yet opened any envelope, this process can be repeated and so again they switch. Now they are back with their original envelopes and yet they think that their fortune has increased 25% twice. By this reasoning, they could convince themselves that by repeatedly switching the envelopes, they could become arbitrarily wealthy. Clearly, something is wrong with the above reasoning, but where is the mistake?

One of the tricks of making paradoxes is to make them slightly more difficult than is necessary to further befuddle us. As John Finn has suggested, in this paradox we could just have well started with a simpler problem. Suppose Ali and Baba know that I am going to give then either an envelope with \$5 or one with \$10 and I am going to toss a coin to decide which to give to Ali, and then give the other to Baba. Then Ali can argue that Baba has $2x$ with probability 1/2 and $x/2$ with probability 1/2. This leads Ali to the same conclusion as before. But now it is clear that this is nonsense, since if Ali has the envelope containing \$5, Baba cannot possibly have half of this, namely \$2.50, since that was not even one of the choices. Similarly, if Ali has \$10, Baba cannot have twice as much, namely \$20. In fact, in this simpler problem the possibly outcomes are given by the tree diagram in Figure 4.14. From the diagram, it is clear that neither is made better off by switching. □

In the above example, Ali's reasoning is incorrect because he infers that if the amount in his envelope is x, then the probability that his envelope contains the

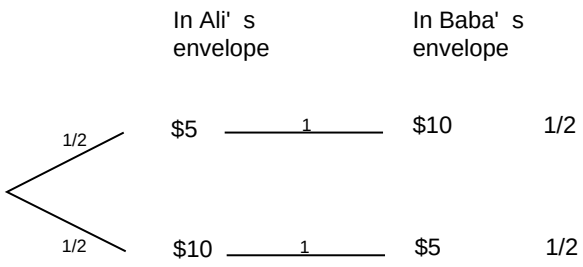

Figure 4.14: John Finn's version of Example 4.28.

smaller amount is 1/2, and the probability that her envelope contains the larger amount is also 1/2. In fact, these conditional probabilities depend upon the distribution of the amounts that are placed in the envelopes.

For definiteness, let X denote the positive integer-valued random variable which represents the smaller of the two amounts in the envelopes. Suppose, in addition, that we are given the distribution of X, i.e., for each positive integer x, we are given the value of

$$p_x = P(X = x) \ .$$

(In Finn's example, $p_5 = 1$, and $p_n = 0$ for all other values of n.) Then it is easy to calculate the conditional probability that an envelope contains the smaller amount, given that it contains x dollars. The two possible sample points are $(x, x/2)$ and $(x, 2x)$. If x is odd, then the first sample point has probability 0, since $x/2$ is not an integer, so the desired conditional probability is 1 that x is the smaller amount. If x is even, then the two sample points have probabilities $p_{x/2}$ and p_x, respectively, so the conditional probability that x is the smaller amount is

$$\frac{p_x}{p_{x/2} + p_x} \ ,$$

which is not necessarily equal to 1/2.

Steven Brams and D. Marc Kilgour[21] study the problem, for different distributions, of whether or not one should switch envelopes, if one's objective is to maximize the long-term average winnings. Let x be the amount in your envelope. They show that for any distribution of X, there is at least one value of x such that you should switch. They give an example of a distribution for which there is exactly one value of x such that you should switch (see Exercise 5). Perhaps the most interesting case is a distribution in which you should always switch. We now give this example.

Example 4.29 Suppose that we have two envelopes in front of us, and that one envelope contains twice the amount of money as the other (both amounts are positive integers). We are given one of the envelopes, and asked if we would like to switch.

[21]S. J. Brams and D. M. Kilgour, "The Box Problem: To Switch or Not to Switch," *Mathematics Magazine*, vol. 68, no. 1 (1995), p. 29.

As above, we let X denote the smaller of the two amounts in the envelopes, and let

$$p_x = P(X = x) \ .$$

We are now in a position where we can calculate the long-term average winnings, if we switch. (This long-term average is an example of a probabilistic concept known as expectation, and will be discussed in Chapter 6.) Given that one of the two sample points has occurred, the probability that it is the point $(x, x/2)$ is

$$\frac{p_{x/2}}{p_{x/2} + p_x} \ ,$$

and the probability that it is the point $(x, 2x)$ is

$$\frac{p_x}{p_{x/2} + p_x} \ .$$

Thus, if we switch, our long-term average winnings are

$$\frac{p_{x/2}}{p_{x/2} + p_x} \frac{x}{2} + \frac{p_x}{p_{x/2} + p_x} 2x \ .$$

If this is greater than x, then it pays in the long run for us to switch. Some routine algebra shows that the above expression is greater than x if and only if

$$\frac{p_{x/2}}{p_{x/2} + p_x} < \frac{2}{3} \ . \tag{4.6}$$

It is interesting to consider whether there is a distribution on the positive integers such that the inequality 4.6 is true for all even values of x. Brams and Kilgour[22] give the following example.

We define p_x as follows:

$$p_x = \begin{cases} \frac{1}{3}\left(\frac{2}{3}\right)^{k-1}, & \text{if } x = 2^k, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to calculate (see Exercise 4) that for all relevant values of x, we have

$$\frac{p_{x/2}}{p_{x/2} + p_x} = \frac{3}{5} \ ,$$

which means that the inequality 4.6 is always true. □

So far, we have been able to resolve paradoxes by clearly stating the assumptions being made and by precisely stating the models being used. We end this section by describing a paradox which we cannot resolve.

Example 4.30 Suppose that we have two envelopes in front of us, and we are told that the envelopes contain X and Y dollars, respectively, where X and Y are different positive integers. We randomly choose one of the envelopes, and we open

[22] ibid.

it, revealing X, say. Is it possible to determine, with probability greater than $1/2$, whether X is the smaller of the two dollar amounts?

Even if we have no knowledge of the joint distribution of X and Y, the surprising answer is yes! Here's how to do it. Toss a fair coin until the first time that heads turns up. Let Z denote the number of tosses required plus $1/2$. If $Z > X$, then we say that X is the smaller of the two amounts, and if $Z < X$, then we say that X is the larger of the two amounts.

First, if Z lies between X and Y, then we are sure to be correct. Since X and Y are unequal, Z lies between them with positive probability. Second, if Z is not between X and Y, then Z is either greater than both X and Y, or is less than both X and Y. In either case, X is the smaller of the two amounts with probability $1/2$, by symmetry considerations (remember, we chose the envelope at random). Thus, the probability that we are correct is greater than $1/2$. □

Exercises

1 One of the first conditional probability paradoxes was provided by Bertrand.[23] It is called the *Box Paradox*. A cabinet has three drawers. In the first drawer there are two gold balls, in the second drawer there are two silver balls, and in the third drawer there is one silver and one gold ball. A drawer is picked at random and a ball chosen at random from the two balls in the drawer. Given that a gold ball was drawn, what is the probability that the drawer with the two gold balls was chosen?

2 The following problem is called the *two aces problem*. This problem, dating back to 1936, has been attributed to the English mathematician J. H. C. Whitehead (see Gridgeman[24]). This problem was also submitted to Marilyn vos Savant by the master of mathematical puzzles Martin Gardner, who remarks that it is one of his favorites.

A bridge hand has been dealt, i. e. thirteen cards are dealt to each player. Given that your partner has at least one ace, what is the probability that he has at least two aces? Given that your partner has the ace of hearts, what is the probability that he has at least two aces? Answer these questions for a version of bridge in which there are eight cards, namely four aces and four kings, and each player is dealt two cards. (The reader may wish to solve the problem with a 52-card deck.)

3 In the preceding exercise, it is natural to ask "How do we get the information that the given hand has an ace?" Gridgeman considers two different ways that we might get this information. (Again, assume the deck consists of eight cards.)

(a) Assume that the person holding the hand is asked to "Name an ace in your hand" and answers "The ace of hearts." What is the probability that he has a second ace?

[23] J. Bertrand, *Calcul des Probabilités*, Gauthier-Uillars, 1888.
[24] N. T. Gridgeman, Letter, *American Statistician*, 21 (1967), pgs. 38-39.

(b) Suppose the person holding the hand is asked the more direct question "Do you have the ace of hearts?" and the answer is yes. What is the probability that he has a second ace?

4 Using the notation introduced in Example 4.29, show that in the example of Brams and Kilgour, if x is a positive power of 2, then

$$\frac{p_{x/2}}{p_{x/2} + p_x} = \frac{3}{5} \ .$$

5 Using the notation introduced in Example 4.29, let

$$p_x = \begin{cases} \frac{2}{3}\left(\frac{1}{3}\right)^k, & \text{if } x = 2^k, \\ 0, & \text{otherwise.} \end{cases}$$

Show that there is exactly one value of x such that if your envelope contains x, then you should switch.

***6** (For bridge players only. From Sutherland.[25]) Suppose that we are the declarer in a hand of bridge, and we have the king, 9, 8, 7, and 2 of a certain suit, while the dummy has the ace, 10, 5, and 4 of the same suit. Suppose that we want to play this suit in such a way as to maximize the probability of having no losers in the suit. We begin by leading the 2 to the ace, and we note that the queen drops on our left. We then lead the 10 from the dummy, and our right-hand opponent plays the six (after playing the three on the first round). Should we finesse or play for the drop?

[25] E. Sutherland, "Restricted Choice — Fact or Fiction?", *Canadian Master Point*, November 1, 1993.

Chapter 5

Important Distributions and Densities

5.1 Important Distributions

In this chapter, we describe the discrete probability distributions and the continuous probability densities that occur most often in the analysis of experiments. We will also show how one simulates these distributions and densities on a computer.

Discrete Uniform Distribution

In Chapter 1, we saw that in many cases, we assume that all outcomes of an experiment are equally likely. If X is a random variable which represents the outcome of an experiment of this type, then we say that X is uniformly distributed. If the sample space S is of size n, where $0 < n < \infty$, then the distribution function $m(\omega)$ is defined to be $1/n$ for all $\omega \in S$. As is the case with all of the discrete probability distributions discussed in this chapter, this experiment can be simulated on a computer using the program **GeneralSimulation**. However, in this case, a faster algorithm can be used instead. (This algorithm was described in Chapter 1; we repeat the description here for completeness.) The expression

$$1 + \lfloor n\,(rnd) \rfloor$$

takes on as a value each integer between 1 and n with probability $1/n$ (the notation $\lfloor x \rfloor$ denotes the greatest integer not exceeding x). Thus, if the possible outcomes of the experiment are labelled $\omega_1\ \omega_2,\ \ldots,\ \omega_n$, then we use the above expression to represent the subscript of the output of the experiment.

If the sample space is a countably infinite set, such as the set of positive integers, then it is not possible to have an experiment which is uniform on this set (see Exercise 3). If the sample space is an uncountable set, with positive, finite length, such as the interval $[0, 1]$, then we use continuous density functions (see Section 5.2).

Binomial Distribution

The binomial distribution with parameters n, p, and k was defined in Chapter 3. It is the distribution of the random variable which counts the number of heads which occur when a coin is tossed n times, assuming that on any one toss, the probability that a head occurs is p. The distribution function is given by the formula

$$b(n,p,k) = \binom{n}{k} p^k q^{n-k} \ ,$$

where $q = 1 - p$.

One straightforward way to simulate a binomial random variable X is to compute the sum of n independent $0-1$ random variables, each of which take on the value 1 with probability p. This method requires n calls to a random number generator to obtain one value of the random variable. When n is relatively large (say at least 30), the Central Limit Theorem (see Chapter 9) implies that the binomial distribution is well-approximated by the corresponding normal density function (which is defined in Section 5.2) with parameters $\mu = np$ and $\sigma = \sqrt{npq}$. Thus, in this case we can compute a value Y of a normal random variable with these parameters, and if $-1/2 \leq Y < n + 1/2$, we can use the value

$$\lfloor Y + 1/2 \rfloor$$

to represent the random variable X. If $Y < -1/2$ or $Y > n + 1/2$, we reject Y and compute another value. We will see in the next section how we can quickly simulate normal random variables.

Geometric Distribution

Consider a Bernoulli trials process continued for an infinite number of trials; for example, a coin tossed an infinite sequence of times. We showed in Section 2.2 how to assign a probability distribution to the infinite tree. Thus, we can determine the distribution for any random variable X relating to the experiment provided $P(X = a)$ can be computed in terms of a finite number of trials. For example, let T be the number of trials up to and including the first success. Then

$$\begin{aligned} P(T = 1) &= p \ , \\ P(T = 2) &= qp \ , \\ P(T = 3) &= q^2 p \ , \end{aligned}$$

and in general,

$$P(T = n) = q^{n-1} p \ .$$

To show that this is a distribution, we must show that

$$p + qp + q^2 p + \cdots = 1 \ .$$

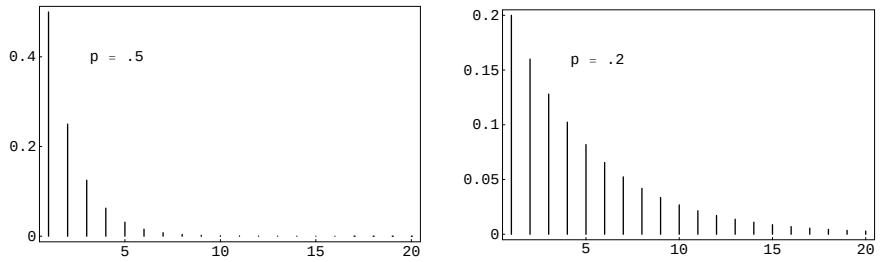

Figure 5.1: Geometric distributions.

The left-hand expression is just a geometric series with first term p and common ratio q, so its sum is

$$\frac{p}{1-q}$$

which equals 1.

In Figure 5.1 we have plotted this distribution using the program **GeometricPlot** for the cases $p = .5$ and $p = .2$. We see that as p decreases we are more likely to get large values for T, as would be expected. In both cases, the most probable value for T is 1. This will always be true since

$$\frac{P(T = j+1)}{P(T = j)} = q < 1 \ .$$

In general, if $0 < p < 1$, and $q = 1 - p$, then we say that the random variable T has a *geometric distribution* if

$$P(T = j) = q^{j-1}p \ ,$$

for $j = 1,\ 2,\ 3,\ \ldots$.

To simulate the geometric distribution with parameter p, we can simply compute a sequence of random numbers in $[0, 1)$, stopping when an entry does not exceed p. However, for small values of p, this is time-consuming (taking, on the average, $1/p$ steps). We now describe a method whose running time does not depend upon the size of p. Define Y to be the smallest integer satisfying the inequality

$$1 - q^Y \geq rnd \ . \tag{5.1}$$

Then we have

$$\begin{aligned} P(Y = j) &= P\Big(1 - q^j \geq rnd > 1 - q^{j-1}\Big) \\ &= q^{j-1} - q^j \\ &= q^{j-1}(1-q) \\ &= q^{j-1}p \ . \end{aligned}$$

Thus, Y is geometrically distributed with parameter p. To generate Y, all we have to do is solve Equation 5.1 for Y. We obtain

$$Y = \left\lceil \frac{\log(1 - rnd)}{\log\ q} \right\rceil ,$$

where the notation $\lceil x \rceil$ means the least integer which is greater than or equal to x. Since $\log(1 - rnd)$ and $\log(rnd)$ are identically distributed, Y can also be generated using the equation

$$Y = \left\lceil \frac{\log\ rnd}{\log\ q} \right\rceil .$$

Example 5.1 The geometric distribution plays an important role in the theory of queues, or waiting lines. For example, suppose a line of customers waits for service at a counter. It is often assumed that, in each small time unit, either 0 or 1 new customers arrive at the counter. The probability that a customer arrives is p and that no customer arrives is $q = 1 - p$. Then the time T until the next arrival has a geometric distribution. It is natural to ask for the probability that no customer arrives in the next k time units, that is, for $P(T > k)$. This is given by

$$\begin{aligned} P(T > k) = \sum_{j=k+1}^{\infty} q^{j-1}p &= q^k(p + qp + q^2p + \cdots) \\ &= q^k \ . \end{aligned}$$

This probability can also be found by noting that we are asking for no successes (i.e., arrivals) in a sequence of k consecutive time units, where the probability of a success in any one time unit is p. Thus, the probability is just q^k, since arrivals in any two time units are independent events.

It is often assumed that the length of time required to service a customer also has a geometric distribution but with a different value for p. This implies a rather special property of the service time. To see this, let us compute the conditional probability

$$P(T > r + s \,|\, T > r) = \frac{P(T > r + s)}{P(T > r)} = \frac{q^{r+s}}{q^r} = q^s \ .$$

Thus, the probability that the customer's service takes s more time units is independent of the length of time r that the customer has already been served. Because of this interpretation, this property is called the "memoryless" property, and is also obeyed by the exponential distribution. (Fortunately, not too many service stations have this property.) □

Negative Binomial Distribution

Suppose we are given a coin which has probability p of coming up heads when it is tossed. We fix a positive integer k, and toss the coin until the kth head appears. We

let X represent the number of tosses. When $k = 1$, X is geometrically distributed. For a general k, we say that X has a negative binomial distribution. We now calculate the probability distribution of X. If $X = x$, then it must be true that there were exactly $k - 1$ heads thrown in the first $x - 1$ tosses, and a head must have been thrown on the xth toss. There are

$$\binom{x-1}{k-1}$$

sequences of length x with these properties, and each of them is assigned the same probability, namely

$$p^{k-1}q^{x-k} \ .$$

Therefore, if we define

$$u(x, k, p) = P(X = x) \ ,$$

then

$$u(x, k, p) = \binom{x-1}{k-1} p^k q^{x-k} \ .$$

One can simulate this on a computer by simulating the tossing of a coin. The following algorithm is, in general, much faster. We note that X can be understood as the sum of k outcomes of a geometrically distributed experiment with parameter p. Thus, we can use the following sum as a means of generating X:

$$\sum_{j=1}^{k} \left\lceil \frac{\log \ rnd_j}{\log \ q} \right\rceil \ .$$

Example 5.2 A fair coin is tossed until the second time a head turns up. The distribution for the number of tosses is $u(x, 2, p)$. Thus the probability that x tosses are needed to obtain two heads is found by letting $k = 2$ in the above formula. We obtain

$$u(x, 2, 1/2) = \binom{x-1}{1} \frac{1}{2^x} \ ,$$

for $x = 2, 3, \ldots$.

In Figure 5.2 we give a graph of the distribution for $k = 2$ and $p = .25$. Note that the distribution is quite asymmetric, with a long tail reflecting the fact that large values of x are possible. □

Poisson Distribution

The Poisson distribution arises in many situations. It is safe to say that it is one of the three most important discrete probability distributions (the other two being the uniform and the binomial distributions). The Poisson distribution can be viewed as arising from the binomial distribution or from the exponential density. We shall now explain its connection with the former; its connection with the latter will be explained in the next section.

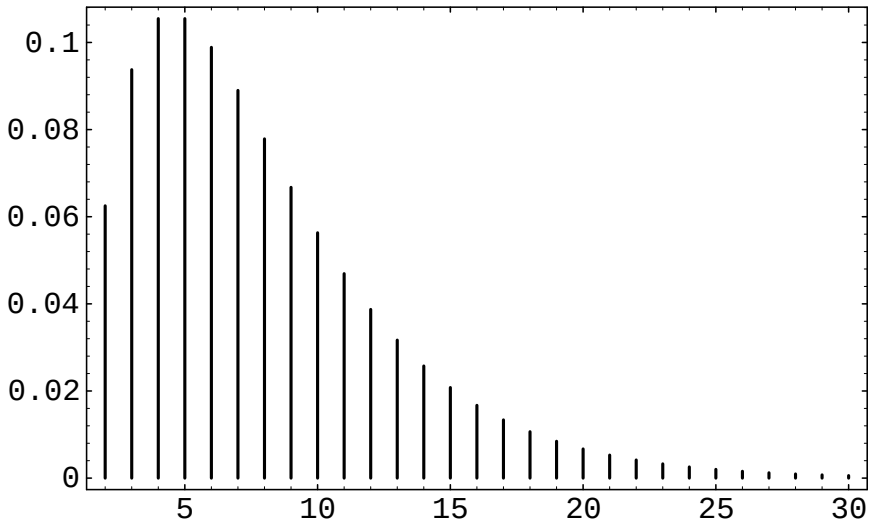

Figure 5.2: Negative binomial distribution with $k = 2$ and $p = .25$.

Suppose that we have a situation in which a certain kind of occurrence happens at random over a period of time. For example, the occurrences that we are interested in might be incoming telephone calls to a police station in a large city. We want to model this situation so that we can consider the probabilities of events such as more than 10 phone calls occurring in a 5-minute time interval. Presumably, in our example, there would be more incoming calls between 6:00 and 7:00 P.M. than between 4:00 and 5:00 A.M., and this fact would certainly affect the above probability. Thus, to have a hope of computing such probabilities, we must assume that the average rate, i.e., the average number of occurrences per minute, is a constant. This rate we will denote by λ. (Thus, in a given 5-minute time interval, we would expect about 5λ occurrences.) This means that if we were to apply our model to the two time periods given above, we would simply use different rates for the two time periods, thereby obtaining two different probabilities for the given event.

Our next assumption is that the number of occurrences in two non-overlapping time intervals are independent. In our example, this means that the events that there are j calls between 5:00 and 5:15 P.M. and k calls between 6:00 and 6:15 P.M. on the same day are independent.

We can use the binomial distribution to model this situation. We imagine that a given time interval is broken up into n subintervals of equal length. If the subintervals are sufficiently short, we can assume that two or more occurrences happen in one subinterval with a probability which is negligible in comparison with the probability of at most one occurrence. Thus, in each subinterval, we are assuming that there is either 0 or 1 occurrence. This means that the sequence of subintervals can be thought of as a sequence of Bernoulli trials, with a success corresponding to an occurrence in the subinterval.

To decide upon the proper value of p, the probability of an occurrence in a given subinterval, we reason as follows. On the average, there are λt occurrences in a time interval of length t. If this time interval is divided into n subintervals, then we would expect, using the Bernoulli trials interpretation, that there should be np occurrences. Thus, we want

$$\lambda t = np \ ,$$

so

$$p = \frac{\lambda t}{n} \ .$$

We now wish to consider the random variable X, which counts the number of occurrences in a given time interval. We want to calculate the distribution of X. For ease of calculation, we will assume that the time interval is of length 1; for time intervals of arbitrary length t, see Exercise 11. We know that

$$P(X = 0) = b(n, p, 0) = (1 - p)^n = \left(1 - \frac{\lambda}{n}\right)^n \ .$$

For large n, this is approximately $e^{-\lambda}$. It is easy to calculate that for any fixed k, we have

$$\frac{b(n, p, k)}{b(n, p, k - 1)} = \frac{\lambda - (k - 1)p}{kq}$$

which, for large n (and therefore small p) is approximately λ/k. Thus, we have

$$P(X = 1) \approx \lambda e^{-\lambda} \ ,$$

and in general,

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda} \ . \tag{5.2}$$

The above distribution is the Poisson distribution. We note that it must be checked that the distribution given in Equation 5.2 really *is* a distribution, i.e., that its values are non-negative and sum to 1. (See Exercise 12.)

The Poisson distribution is used as an approximation to the binomial distribution when the parameters n and p are large and small, respectively (see Examples 5.3 and 5.4). However, the Poisson distribution also arises in situations where it may not be easy to interpret or measure the parameters n and p (see Example 5.5).

Example 5.3 A typesetter makes, on the average, one mistake per 1000 words. Assume that he is setting a book with 100 words to a page. Let S_{100} be the number of mistakes that he makes on a single page. Then the exact probability distribution for S_{100} would be obtained by considering S_{100} as a result of 100 Bernoulli trials with $p = 1/1000$. The expected value of S_{100} is $\lambda = 100(1/1000) = .1$. The exact probability that $S_{100} = j$ is $b(100, 1/1000, j)$, and the Poisson approximation is

$$\frac{e^{-.1}(.1)^j}{j!}.$$

In Table 5.1 we give, for various values of n and p, the exact values computed by the binomial distribution and the Poisson approximation. □

j	Poisson $\lambda = .1$	Binomial $n = 100$ $p = .001$	Poisson $\lambda = 1$	Binomial $n = 100$ $p = .01$	Poisson $\lambda = 10$	Binomial $n = 1000$ $p = .01$
0	.9048	.9048	.3679	.3660	.0000	.0000
1	.0905	.0905	.3679	.3697	.0005	.0004
2	.0045	.0045	.1839	.1849	.0023	.0022
3	.0002	.0002	.0613	.0610	.0076	.0074
4	.0000	.0000	.0153	.0149	.0189	.0186
5			.0031	.0029	.0378	.0374
6			.0005	.0005	.0631	.0627
7			.0001	.0001	.0901	.0900
8			.0000	.0000	.1126	.1128
9					.1251	.1256
10					.1251	.1257
11					.1137	.1143
12					.0948	.0952
13					.0729	.0731
14					.0521	.0520
15					.0347	.0345
16					.0217	.0215
17					.0128	.0126
18					.0071	.0069
19					.0037	.0036
20					.0019	.0018
21					.0009	.0009
22					.0004	.0004
23					.0002	.0002
24					.0001	.0001
25					.0000	.0000

Table 5.1: Poisson approximation to the binomial distribution.

Example 5.4 In his book,[1] Feller discusses the statistics of flying bomb hits in the south of London during the Second World War.

Assume that you live in a district of size 10 blocks by 10 blocks so that the total district is divided into 100 small squares. How likely is it that the square in which you live will receive no hits if the total area is hit by 400 bombs?

We assume that a particular bomb will hit your square with probability 1/100. Since there are 400 bombs, we can regard the number of hits that your square receives as the number of *successes* in a Bernoulli trials process with $n = 400$ and $p = 1/100$. Thus we can use the Poisson distribution with $\lambda = 400 \cdot 1/100 = 4$ to approximate the probability that your square will receive j hits. This probability is $p(j) = e^{-4}4^j/j!$. The expected number of squares that receive exactly j hits is then $100 \cdot p(j)$. It is easy to write a program **LondonBombs** to simulate this situation and compare the expected number of squares with j hits with the observed number. In Exercise 26 you are asked to compare the actual observed data with that predicted by the Poisson distribution.

In Figure 5.3, we have shown the simulated hits, together with a spike graph showing both the observed and predicted frequencies. The observed frequencies are shown as squares, and the predicted frequencies are shown as dots. □

If the reader would rather not consider flying bombs, he is invited to instead consider an analogous situation involving cookies and raisins. We assume that we have made enough cookie dough for 500 cookies. We put 600 raisins in the dough, and mix it thoroughly. One way to look at this situation is that we have 500 cookies, and after placing the cookies in a grid on the table, we throw 600 raisins at the cookies. (See Exercise 22.)

Example 5.5 Suppose that in a certain fixed amount A of blood, the average human has 40 white blood cells. Let X be the random variable which gives the number of white blood cells in a random sample of size A from a random individual. We can think of X as binomially distributed with each white blood cell in the body representing a trial. If a given white blood cell turns up in the sample, then the trial corresponding to that blood cell was a success. Then p should be taken as the ratio of A to the total amount of blood in the individual, and n will be the number of white blood cells in the individual. Of course, in practice, neither of these parameters is very easy to measure accurately, but presumably the number 40 is easy to measure. But for the average human, we then have $40 = np$, so we can think of X as being Poisson distributed, with parameter $\lambda = 40$. In this case, it is easier to model the situation using the Poisson distribution than the binomial distribution. □

To simulate a Poisson random variable on a computer, a good way is to take advantage of the relationship between the Poisson distribution and the exponential density. This relationship and the resulting simulation algorithm will be described in the next section.

[1] ibid., p. 161.

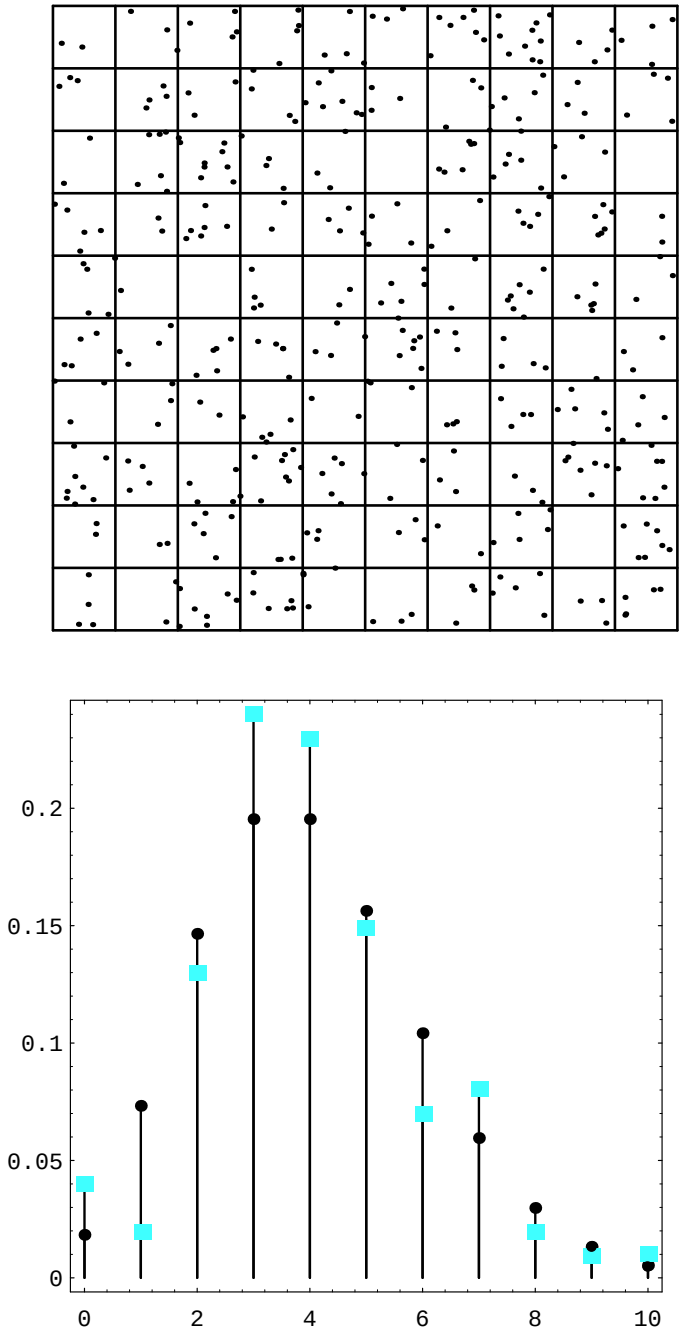

Figure 5.3: Flying bomb hits.

Hypergeometric Distribution

Suppose that we have a set of N balls, of which k are red and $N-k$ are blue. We choose n of these balls, without replacement, and define X to be the number of red balls in our sample. The distribution of X is called the hypergeometric distribution. We note that this distribution depends upon three parameters, namely N, k, and n. There does not seem to be a standard notation for this distribution; we will use the notation $h(N,k,n,x)$ to denote $P(X=x)$. This probability can be found by noting that there are

$$\binom{N}{n}$$

different samples of size n, and the number of such samples with exactly x red balls is obtained by multiplying the number of ways of choosing x red balls from the set of k red balls and the number of ways of choosing $n-x$ blue balls from the set of $N-k$ blue balls. Hence, we have

$$h(N,k,n,x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} .$$

This distribution can be generalized to the case where there are more than two types of objects. (See Exercise 40.)

If we let N and k tend to ∞, in such a way that the ratio k/N remains fixed, then the hypergeometric distribution tends to the binomial distribution with parameters n and $p = k/N$. This is reasonable because if N and k are much larger than n, then whether we choose our sample with or without replacement should not affect the probabilities very much, and the experiment consisting of choosing with replacement yields a binomially distributed random variable (see Exercise 44).

An example of how this distribution might be used is given in Exercises 36 and 37. We now give another example involving the hypergeometric distribution. It illustrates a statistical test called Fisher's Exact Test.

Example 5.6 It is often of interest to consider two traits, such as eye color and hair color, and to ask whether there is an association between the two traits. Two traits are associated if knowing the value of one of the traits for a given person allows us to predict the value of the other trait for that person. The stronger the association, the more accurate the predictions become. If there is no association between the traits, then we say that the traits are independent. In this example, we will use the traits of gender and political party, and we will assume that there are only two possible genders, female and male, and only two possible political parties, Democratic and Republican.

Suppose that we have collected data concerning these traits. To test whether there is an association between the traits, we first assume that there is no association between the two traits. This gives rise to an "expected" data set, in which knowledge of the value of one trait is of no help in predicting the value of the other trait. Our collected data set usually differs from this expected data set. If it differs by quite a bit, then we would tend to reject the assumption of independence of the traits. To

	Democrat	Republican	
Female	24	4	28
Male	8	14	22
	32	18	50

Table 5.2: Observed data.

	Democrat	Republican	
Female	s_{11}	s_{12}	t_{11}
Male	s_{21}	s_{22}	t_{12}
	t_{21}	t_{22}	n

Table 5.3: General data table.

nail down what is meant by "quite a bit," we decide which possible data sets differ from the expected data set by at least as much as ours does, and then we compute the probability that any of these data sets would occur under the assumption of independence of traits. If this probability is small, then it is unlikely that the difference between our collected data set and the expected data set is due entirely to chance.

Suppose that we have collected the data shown in Table 5.2. The row and column sums are called marginal totals, or marginals. In what follows, we will denote the row sums by t_{11} and t_{12}, and the column sums by t_{21} and t_{22}. The ijth entry in the table will be denoted by s_{ij}. Finally, the size of the data set will be denoted by n. Thus, a general data table will look as shown in Table 5.3. We now explain the model which will be used to construct the "expected" data set. In the model, we assume that the two traits are independent. We then put t_{21} yellow balls and t_{22} green balls, corresponding to the Democratic and Republican marginals, into an urn. We draw t_{11} balls, without replacement, from the urn, and call these balls females. The t_{12} balls remaining in the urn are called males. In the specific case under consideration, the probability of getting the actual data under this model is given by the expression

$$\frac{\binom{32}{24}\binom{18}{4}}{\binom{50}{28}},$$

i.e., a value of the hypergeometric distribution.

We are now ready to construct the expected data set. If we choose 28 balls out of 50, we should expect to see, on the average, the same percentage of yellow balls in our sample as in the urn. Thus, we should expect to see, on the average, $28(32/50) = 17.92 \approx 18$ yellow balls in our sample. (See Exercise 36.) The other expected values are computed in exactly the same way. Thus, the expected data set is shown in Table 5.4. We note that the value of s_{11} determines the other three values in the table, since the marginals are all fixed. Thus, in considering the possible data sets that could appear in this model, it is enough to consider the various possible values of s_{11}. In the specific case at hand, what is the probability

	Democrat	Republican	
Female	18	10	28
Male	14	8	22
	32	18	50

Table 5.4: Expected data.

of drawing exactly a yellow balls, i.e., what is the probability that $s_{11} = a$? It is

$$\frac{\binom{32}{a}\binom{18}{28-a}}{\binom{50}{28}} \ . \tag{5.3}$$

We are now ready to decide whether our actual data differs from the expected data set by an amount which is greater than could be reasonably attributed to chance alone. We note that the expected number of female Democrats is 18, but the actual number in our data is 24. The other data sets which differ from the expected data set by more than ours correspond to those where the number of female Democrats equals 25, 26, 27, or 28. Thus, to obtain the required probability, we sum the expression in (5.3) from $a = 24$ to $a = 28$. We obtain a value of .000395. Thus, we should reject the hypothesis that the two traits are independent. □

Finally, we turn to the question of how to simulate a hypergeometric random variable X. Let us assume that the parameters for X are N, k, and n. We imagine that we have a set of N balls, labelled from 1 to N. We decree that the first k of these balls are red, and the rest are blue. Suppose that we have chosen m balls, and that j of them are red. Then there are $k - j$ red balls left, and $N - m$ balls left. Thus, our next choice will be red with probability

$$\frac{k-j}{N-m} \ .$$

So at this stage, we choose a random number in $[0, 1]$, and report that a red ball has been chosen if and only if the random number does not exceed the above expression. Then we update the values of m and j, and continue until n balls have been chosen.

Benford Distribution

Our next example of a distribution comes from the study of leading digits in data sets. It turns out that many data sets that occur "in real life" have the property that the first digits of the data are not uniformly distributed over the set $\{1, 2, \ldots, 9\}$. Rather, it appears that the digit 1 is most likely to occur, and that the distribution is monotonically decreasing on the set of possible digits. The Benford distribution appears, in many cases, to fit such data. Many explanations have been given for the occurrence of this distribution. Possibly the most convincing explanation is that this distribution is the only one that is invariant under a change of scale. If one thinks of certain data sets as somehow "naturally occurring," then the distribution should be unaffected by which units are chosen in which to represent the data, i.e., the distribution should be invariant under change of scale.

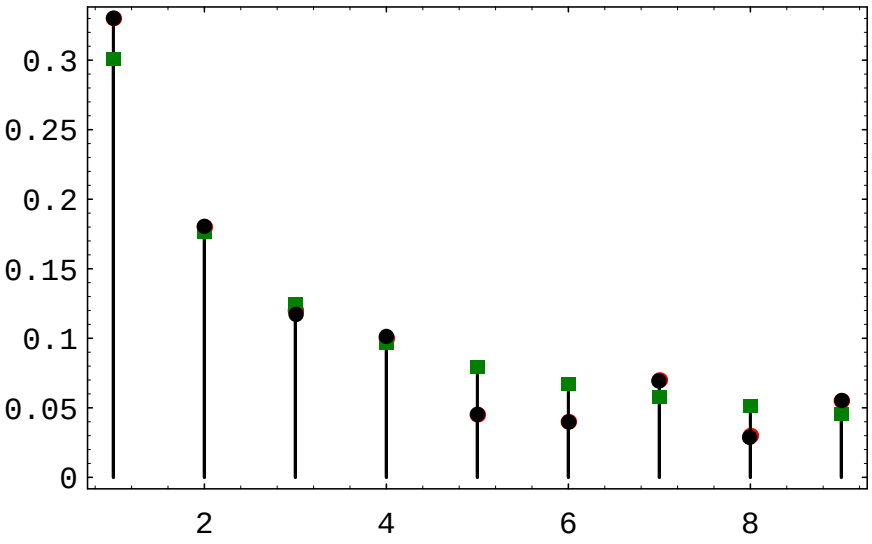

Figure 5.4: Leading digits in President Clinton's tax returns.

Theodore Hill[2] gives a general description of the Benford distribution, when one considers the first d digits of integers in a data set. We will restrict our attention to the first digit. In this case, the Benford distribution has distribution function

$$f(k) = \log_{10}(k+1) - \log_{10}(k) \ ,$$

for $1 \leq k \leq 9$.

Mark Nigrini[3] has advocated the use of the Benford distribution as a means of testing suspicious financial records such as bookkeeping entries, checks, and tax returns. His idea is that if someone were to "make up" numbers in these cases, the person would probably produce numbers that are fairly uniformly distributed, while if one were to use the actual numbers, the leading digits would roughly follow the Benford distribution. As an example, Nigrini analyzed President Clinton's tax returns for a 13-year period. In Figure 5.4, the Benford distribution values are shown as squares, and the President's tax return data are shown as circles. One sees that in this example, the Benford distribution fits the data very well.

This distribution was discovered by the astronomer Simon Newcomb who stated the following in his paper on the subject: "That the ten digits do not occur with equal frequency must be evident to anyone making use of logarithm tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9."[4]

[2]T. P. Hill, "The Significant Digit Phenomenon," *American Mathematical Monthly,* vol. 102, no. 4 (April 1995), pgs. 322-327.

[3]M. Nigrini, "Detecting Biases and Irregularities in Tabulated Data," working paper

[4]S. Newcomb, "Note on the frequency of use of the different digits in natural numbers," *American Journal of Mathematics,* vol. 4 (1881), pgs. 39-40.

Exercises

1 For which of the following random variables would it be appropriate to assign a uniform distribution?

(a) Let X represent the roll of one die.

(b) Let X represent the number of heads obtained in three tosses of a coin.

(c) A roulette wheel has 38 possible outcomes: 0, 00, and 1 through 36. Let X represent the outcome when a roulette wheel is spun.

(d) Let X represent the birthday of a randomly chosen person.

(e) Let X represent the number of tosses of a coin necessary to achieve a head for the first time.

2 Let n be a positive integer. Let S be the set of integers between 1 and n. Consider the following process: We remove a number from S at random and write it down. We repeat this until S is empty. The result is a permutation of the integers from 1 to n. Let X denote this permutation. Is X uniformly distributed?

3 Let X be a random variable which can take on countably many values. Show that X cannot be uniformly distributed.

4 Suppose we are attending a college which has 3000 students. We wish to choose a subset of size 100 from the student body. Let X represent the subset, chosen using the following possible strategies. For which strategies would it be appropriate to assign the uniform distribution to X? If it is appropriate, what probability should we assign to each outcome?

(a) Take the first 100 students who enter the cafeteria to eat lunch.

(b) Ask the Registrar to sort the students by their Social Security number, and then take the first 100 in the resulting list.

(c) Ask the Registrar for a set of cards, with each card containing the name of exactly one student, and with each student appearing on exactly one card. Throw the cards out of a third-story window, then walk outside and pick up the first 100 cards that you find.

5 Under the same conditions as in the preceding exercise, can you describe a procedure which, if used, would produce each possible outcome with the same probability? Can you describe such a procedure that does not rely on a computer or a calculator?

6 Let $X_1,\ X_2,\ \ldots,\ X_n$ be n mutually independent random variables, each of which is uniformly distributed on the integers from 1 to k. Let Y denote the minimum of the X_i's. Find the distribution of Y.

7 A die is rolled until the first time T that a six turns up.

(a) What is the probability distribution for T?

(b) Find $P(T > 3)$.

(c) Find $P(T > 6|T > 3)$.

8 If a coin is tossed a sequence of times, what is the probability that the first head will occur after the fifth toss, given that it has not occurred in the first two tosses?

9 A worker for the Department of Fish and Game is assigned the job of estimating the number of trout in a certain lake of modest size. She proceeds as follows: She catches 100 trout, tags each of them, and puts them back in the lake. One month later, she catches 100 more trout, and notes that 10 of them have tags.

(a) Without doing any fancy calculations, give a rough estimate of the number of trout in the lake.

(b) Let N be the number of trout in the lake. Find an expression, in terms of N, for the probability that the worker would catch 10 tagged trout out of the 100 trout that she caught the second time.

(c) Find the value of N which maximizes the expression in part (b). This value is called the *maximum likelihood estimate* for the unknown quantity N. *Hint*: Consider the ratio of the expressions for successive values of N.

10 A census in the United States is an attempt to count everyone in the country. It is inevitable that many people are not counted. The U. S. Census Bureau proposed a way to estimate the number of people who were not counted by the latest census. Their proposal was as follows: In a given locality, let N denote the actual number of people who live there. Assume that the census counted n_1 people living in this area. Now, another census was taken in the locality, and n_2 people were counted. In addition, n_{12} people were counted both times.

(a) Given N, n_1, and n_2, let X denote the number of people counted both times. Find the probability that $X = k$, where k is a fixed positive integer between 0 and n_2.

(b) Now assume that $X = n_{12}$. Find the value of N which maximizes the expression in part (a). *Hint*: Consider the ratio of the expressions for successive values of N.

11 Suppose that X is a random variable which represents the number of calls coming in to a police station in a one-minute interval. In the text, we showed that X could be modelled using a Poisson distribution with parameter λ, where this parameter represents the average number of incoming calls per minute. Now suppose that Y is a random variable which represents the number of incoming calls in an interval of length t. Show that the distribution of Y is given by

$$P(Y = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!} \ ,$$

i.e., Y is Poisson with parameter λt. *Hint*: Suppose a Martian were to observe the police station. Let us also assume that the basic time interval used on Mars is exactly t Earth minutes. Finally, we will assume that the Martian understands the derivation of the Poisson distribution in the text. What would she write down for the distribution of Y?

12 Show that the values of the Poisson distribution given in Equation 5.2 sum to 1.

13 The Poisson distribution with parameter $\lambda = .3$ has been assigned for the outcome of an experiment. Let X be the outcome function. Find $P(X = 0)$, $P(X = 1)$, and $P(X > 1)$.

14 On the average, only 1 person in 1000 has a particular rare blood type.

(a) Find the probability that, in a city of 10,000 people, no one has this blood type.

(b) How many people would have to be tested to give a probability greater than 1/2 of finding at least one person with this blood type?

15 Write a program for the user to input n, p, j and have the program print out the exact value of $b(n, p, k)$ and the Poisson approximation to this value.

16 Assume that, during each second, a Dartmouth switchboard receives one call with probability .01 and no calls with probability .99. Use the Poisson approximation to estimate the probability that the operator will miss at most one call if she takes a 5-minute coffee break.

17 The probability of a royal flush in a poker hand is $p = 1/649{,}740$. How large must n be to render the probability of having no royal flush in n hands smaller than $1/e$?

18 A baker blends 600 raisins and 400 chocolate chips into a dough mix and, from this, makes 500 cookies.

(a) Find the probability that a randomly picked cookie will have no raisins.

(b) Find the probability that a randomly picked cookie will have exactly two chocolate chips.

(c) Find the probability that a randomly chosen cookie will have at least two bits (raisins or chips) in it.

19 The probability that, in a bridge deal, one of the four hands has all hearts is approximately 6.3×10^{-12}. In a city with about 50,000 bridge players the resident probability expert is called on the average once a year (usually late at night) and told that the caller has just been dealt a hand of all hearts. Should she suspect that some of these callers are the victims of practical jokes?

20 An advertiser drops 10,000 leaflets on a city which has 2000 blocks. Assume that each leaflet has an equal chance of landing on each block. What is the probability that a particular block will receive no leaflets?

21 In a class of 80 students, the professor calls on 1 student chosen at random for a recitation in each class period. There are 32 class periods in a term.

(a) Write a formula for the exact probability that a given student is called upon j times during the term.

(b) Write a formula for the Poisson approximation for this probability. Using your formula estimate the probability that a given student is called upon more than twice.

22 Assume that we are making raisin cookies. We put a box of 600 raisins into our dough mix, mix up the dough, then make from the dough 500 cookies. We then ask for the probability that a randomly chosen cookie will have 0, 1, 2, ... raisins. Consider the cookies as trials in an experiment, and let X be the random variable which gives the number of raisins in a given cookie. Then we can regard the number of raisins in a cookie as the result of $n = 600$ independent trials with probability $p = 1/500$ for success on each trial. Since n is large and p is small, we can use the Poisson approximation with $\lambda = 600(1/500) = 1.2$. Determine the probability that a given cookie will have at least five raisins.

23 For a certain experiment, the Poisson distribution with parameter $\lambda = m$ has been assigned. Show that a most probable outcome for the experiment is the integer value k such that $m - 1 \leq k \leq m$. Under what conditions will there be two most probable values? *Hint*: Consider the ratio of successive probabilities.

24 When John Kemeny was chair of the Mathematics Department at Dartmouth College, he received an average of ten letters each day. On a certain weekday he received no mail and wondered if it was a holiday. To decide this he computed the probability that, in ten years, he would have at least 1 day without any mail. He assumed that the number of letters he received on a given day has a Poisson distribution. What probability did he find? *Hint*: Apply the Poisson distribution twice. First, to find the probability that, in 3000 days, he will have at least 1 day without mail, assuming each year has about 300 days on which mail is delivered.

25 Reese Prosser never puts money in a 10-cent parking meter in Hanover. He assumes that there is a probability of .05 that he will be caught. The first offense costs nothing, the second costs 2 dollars, and subsequent offenses cost 5 dollars each. Under his assumptions, how does the expected cost of parking 100 times without paying the meter compare with the cost of paying the meter each time?

Number of deaths	Number of corps with x deaths in a given year
0	144
1	91
2	32
3	11
4	2

Table 5.5: Mule kicks.

26 Feller[5] discusses the statistics of flying bomb hits in an area in the south of London during the Second World War. The area in question was divided into $24 \times 24 = 576$ small areas. The total number of hits was 537. There were 229 squares with 0 hits, 211 with 1 hit, 93 with 2 hits, 35 with 3 hits, 7 with 4 hits, and 1 with 5 or more. Assuming the hits were purely random, use the Poisson approximation to find the probability that a particular square would have exactly k hits. Compute the expected number of squares that would have 0, 1, 2, 3, 4, and 5 or more hits and compare this with the observed results.

27 Assume that the probability that there is a significant accident in a nuclear power plant during one year's time is .001. If a country has 100 nuclear plants, estimate the probability that there is at least one such accident during a given year.

28 An airline finds that 4 percent of the passengers that make reservations on a particular flight will not show up. Consequently, their policy is to sell 100 reserved seats on a plane that has only 98 seats. Find the probability that every person who shows up for the flight will find a seat available.

29 The king's coinmaster boxes his coins 500 to a box and puts 1 counterfeit coin in each box. The king is suspicious, but, instead of testing all the coins in 1 box, he tests 1 coin chosen at random out of each of 500 boxes. What is the probability that he finds at least one fake? What is it if the king tests 2 coins from each of 250 boxes?

30 (From Kemeny[6]) Show that, if you make 100 bets on the number 17 at roulette at Monte Carlo (see Example 6.13), you will have a probability greater than 1/2 of coming out ahead. What is your expected winning?

31 In one of the first studies of the Poisson distribution, von Bortkiewicz[7] considered the frequency of deaths from kicks in the Prussian army corps. From the study of 14 corps over a 20-year period, he obtained the data shown in Table 5.5. Fit a Poisson distribution to this data and see if you think that the Poisson distribution is appropriate.

[5] ibid., p. 161.
[6] Private communication.
[7] L. von Bortkiewicz, *Das Gesetz der Kleinen Zahlen* (Leipzig: Teubner, 1898), p. 24.

32 It is often assumed that the auto traffic that arrives at the intersection during a unit time period has a Poisson distribution with expected value m. Assume that the number of cars X that arrive at an intersection from the north in unit time has a Poisson distribution with parameter $\lambda = m$ and the number Y that arrive from the west in unit time has a Poisson distribution with parameter $\lambda = \bar{m}$. If X and Y are independent, show that the total number $X + Y$ that arrive at the intersection in unit time has a Poisson distribution with parameter $\lambda = m + \bar{m}$.

33 Cars coming along Magnolia Street come to a fork in the road and have to choose either Willow Street or Main Street to continue. Assume that the number of cars that arrive at the fork in unit time has a Poisson distribution with parameter $\lambda = 4$. A car arriving at the fork chooses Main Street with probability 3/4 and Willow Street with probability 1/4. Let X be the random variable which counts the number of cars that, in a given unit of time, pass by Joe's Barber Shop on Main Street. What is the distribution of X?

34 In the appeal of the *People v. Collins* case (see Exercise 4.1.28), the counsel for the defense argued as follows: Suppose, for example, there are 5,000,000 couples in the Los Angeles area and the probability that a randomly chosen couple fits the witnesses' description is 1/12,000,000. Then the probability that there are two such couples given that there is at least one is not at all small. Find this probability. (The California Supreme Court overturned the initial guilty verdict.)

35 A manufactured lot of brass turnbuckles has S items of which D are defective. A sample of s items is drawn without replacement. Let X be a random variable that gives the number of defective items in the sample. Let $p(d) = P(X = d)$.

(a) Show that

$$p(d) = \frac{\binom{D}{d}\binom{S-D}{s-d}}{\binom{S}{s}} .$$

Thus, X is hypergeometric.

(b) Prove the following identity, known as *Euler's formula*:

$$\sum_{d=0}^{\min(D,s)} \binom{D}{d}\binom{S-D}{s-d} = \binom{S}{s} .$$

36 A bin of 1000 turnbuckles has an unknown number D of defectives. A sample of 100 turnbuckles has 2 defectives. The *maximum likelihood estimate* for D is the number of defectives which gives the highest probability for obtaining the number of defectives observed in the sample. Guess this number D and then write a computer program to verify your guess.

37 There are an unknown number of moose on Isle Royale (a National Park in Lake Superior). To estimate the number of moose, 50 moose are captured and

tagged. Six months later 200 moose are captured and it is found that 8 of these were tagged. Estimate the number of moose on Isle Royale from these data, and then verify your guess by computer program (see Exercise 36).

38 A manufactured lot of buggy whips has 20 items, of which 5 are defective. A random sample of 5 items is chosen to be inspected. Find the probability that the sample contains exactly one defective item

(a) if the sampling is done with replacement.

(b) if the sampling is done without replacement.

39 Suppose that N and k tend to ∞ in such a way that k/N remains fixed. Show that

$$h(N, k, n, x) \to b(n, k/N, x) \ .$$

40 A bridge deck has 52 cards with 13 cards in each of four suits: spades, hearts, diamonds, and clubs. A hand of 13 cards is dealt from a shuffled deck. Find the probability that the hand has

(a) a distribution of suits 4, 4, 3, 2 (for example, four spades, four hearts, three diamonds, two clubs).

(b) a distribution of suits 5, 3, 3, 2.

41 Write a computer algorithm that simulates a hypergeometric random variable with parameters N, k, and n.

42 You are presented with four different dice. The first one has two sides marked 0 and four sides marked 4. The second one has a 3 on every side. The third one has a 2 on four sides and a 6 on two sides, and the fourth one has a 1 on three sides and a 5 on three sides. You allow your friend to pick any of the four dice he wishes. Then you pick one of the remaining three and you each roll your die. The person with the largest number showing wins a dollar. Show that you can choose your die so that you have probability 2/3 of winning no matter which die your friend picks. (See Tenney and Foster.[8])

43 The students in a certain class were classified by hair color and eye color. The conventions used were: Brown and black hair were considered dark, and red and blonde hair were considered light; black and brown eyes were considered dark, and blue and green eyes were considered light. They collected the data shown in Table 5.6. Are these traits independent? (See Example 5.6.)

44 Suppose that in the hypergeometric distribution, we let N and k tend to ∞ in such a way that the ratio k/N approaches a real number p between 0 and 1. Show that the hypergeometric distribution tends to the binomial distribution with parameters n and p.

[8] R. L. Tenney and C. C. Foster, *Non-transitive Dominance*, Math. Mag. 49 (1976) no. 3, pgs. 115-120.

	Dark Eyes	Light Eyes	
Dark Hair	28	15	43
Light Hair	9	23	32
	37	38	75

Table 5.6: Observed data.

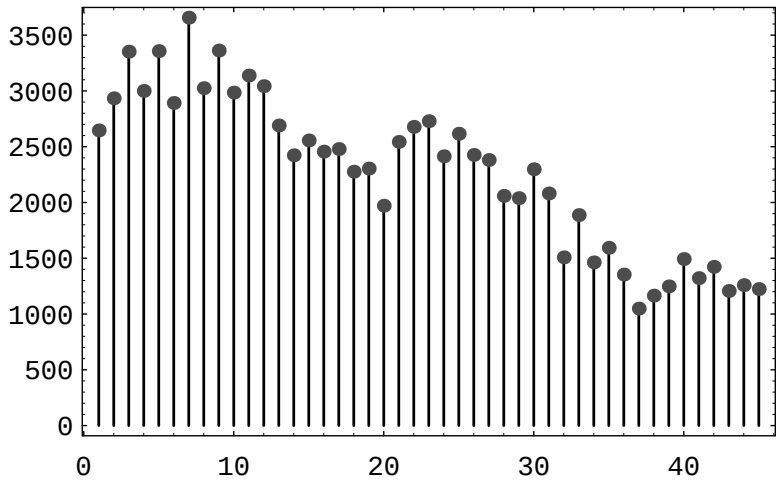

Figure 5.5: Distribution of choices in the Powerball lottery.

45 (a) Compute the leading digits of the first 100 powers of 2, and see how well these data fit the Benford distribution.

(b) Multiply each number in the data set of part (a) by 3, and compare the distribution of the leading digits with the Benford distribution.

46 In the Powerball lottery, contestants pick 5 different integers between 1 and 45, and in addition, pick a bonus integer from the same range (the bonus integer can equal one of the first five integers chosen). Some contestants choose the numbers themselves, and others let the computer choose the numbers. The data shown in Table 5.7 are the contestant-chosen numbers in a certain state on May 3, 1996. A spike graph of the data is shown in Figure 5.5.

The goal of this problem is to check the hypothesis that the chosen numbers are uniformly distributed. To do this, compute the value v of the random variable χ^2 given in Example 5.6. In the present case, this random variable has 44 degrees of freedom. One can find, in a χ^2 table, the value $v_0 = 59.43$, which represents a number with the property that a χ^2-distributed random variable takes on values that exceed v_0 only 5% of the time. Does your computed value of v exceed v_0? If so, you should reject the hypothesis that the contestants' choices are uniformly distributed.

Integer	Times Chosen	Integer	Times Chosen	Integer	Times Chosen
1	2646	2	2934	3	3352
4	3000	5	3357	6	2892
7	3657	8	3025	9	3362
10	2985	11	3138	12	3043
13	2690	14	2423	15	2556
16	2456	17	2479	18	2276
19	2304	20	1971	21	2543
22	2678	23	2729	24	2414
25	2616	26	2426	27	2381
28	2059	29	2039	30	2298
31	2081	32	1508	33	1887
34	1463	35	1594	36	1354
37	1049	38	1165	39	1248
40	1493	41	1322	42	1423
43	1207	44	1259	45	1224

Table 5.7: Numbers chosen by contestants in the Powerball lottery.

5.2 Important Densities

In this section, we will introduce some important probability density functions and give some examples of their use. We will also consider the question of how one simulates a given density using a computer.

Continuous Uniform Density

The simplest density function corresponds to the random variable U whose value represents the outcome of the experiment consisting of choosing a real number at random from the interval $[a, b]$.

$$f(\omega) = \begin{cases} 1/(b-a), & \text{if } a \leq \omega \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to simulate this density on a computer. We simply calculate the expression

$$(b-a)rnd + a \ .$$

Exponential and Gamma Densities

The exponential density function is defined by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } 0 \leq x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Here λ is any positive constant, depending on the experiment. The reader has seen this density in Example 2.17. In Figure 5.6 we show graphs of several exponential densities for different choices of λ. The exponential density is often used to

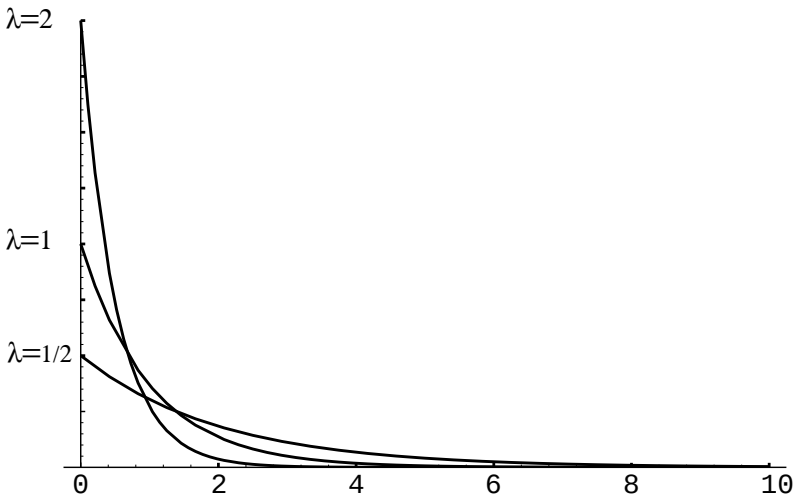

Figure 5.6: Exponential densities.

describe experiments involving a question of the form: How long until something happens? For example, the exponential density is often used to study the time between emissions of particles from a radioactive source.

The cumulative distribution function of the exponential density is easy to compute. Let T be an exponentially distributed random variable with parameter λ. If $x \geq 0$, then we have

$$\begin{aligned} F(x) &= P(T \leq x) \\ &= \int_0^x \lambda e^{-\lambda t}\, dt \\ &= 1 - e^{-\lambda x} \ . \end{aligned}$$

Both the exponential density and the geometric distribution share a property known as the "memoryless" property. This property was introduced in Example 5.1; it says that

$$P(T > r + s \,|\, T > r) = P(T > s) \ .$$

This can be demonstrated to hold for the exponential density by computing both sides of this equation. The right-hand side is just

$$1 - F(s) = e^{-\lambda s} \ ,$$

while the left-hand side is

$$\frac{P(T > r + s)}{P(T > r)} = \frac{1 - F(r + s)}{1 - F(r)}$$

$$\begin{aligned} &= \frac{e^{-\lambda(r+s)}}{e^{-\lambda r}} \\ &= e^{-\lambda s} \ . \end{aligned}$$

There is a very important relationship between the exponential density and the Poisson distribution. We begin by defining $X_1,\ X_2,\ \ldots$ to be a sequence of independent exponentially distributed random variables with parameter λ. We might think of X_i as denoting the amount of time between the ith and $(i+1)$st emissions of a particle by a radioactive source. (As we shall see in Chapter 6, we can think of the parameter λ as representing the reciprocal of the average length of time between emissions. This parameter is a quantity that might be measured in an actual experiment of this type.)

We now consider a time interval of length t, and we let Y denote the random variable which counts the number of emissions that occur in the time interval. We would like to calculate the distribution function of Y (clearly, Y is a discrete random variable). If we let S_n denote the sum $X_1 + X_2 + \cdots + X_n$, then it is easy to see that

$$P(Y = n) = P(S_n \leq t \text{ and } S_{n+1} > t) \ .$$

Since the event $S_{n+1} \leq t$ is a subset of the event $S_n \leq t$, the above probability is seen to be equal to

$$P(S_n \leq t) - P(S_{n+1} \leq t) \ . \tag{5.4}$$

We will show in Chapter 7 that the density of S_n is given by the following formula:

$$g_n(x) = \begin{cases} \lambda \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This density is an example of a gamma density with parameters λ and n. The general gamma density allows n to be any positive real number. We shall not discuss this general density.

It is easy to show by induction on n that the cumulative distribution function of S_n is given by:

$$G_n(x) = \begin{cases} 1 - e^{-\lambda x}\left(1 + \frac{\lambda x}{1!} + \cdots + \frac{(\lambda x)^{n-1}}{(n-1)!}\right), & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Using this expression, the quantity in (5.4) is easy to compute; we obtain

$$e^{-\lambda t}\frac{(\lambda t)^n}{n!} \ ,$$

which the reader will recognize as the probability that a Poisson-distributed random variable, with parameter λt, takes on the value n.

The above relationship will allow us to simulate a Poisson distribution, once we have found a way to simulate an exponential density. The following random variable does the job:

$$Y = -\frac{1}{\lambda}\log(rnd) \ . \tag{5.5}$$

Using Corollary 5.2 (below), one can derive the above expression (see Exercise 3). We content ourselves for now with a short calculation that should convince the reader that the random variable Y has the required property. We have

$$\begin{aligned} P(Y \leq y) &= P\Big(-\frac{1}{\lambda}\log(rnd) \leq y\Big) \\ &= P(\log(rnd) \geq -\lambda y) \\ &= P(rnd \geq e^{-\lambda y}) \\ &= 1 - e^{-\lambda y} \ . \end{aligned}$$

This last expression is seen to be the cumulative distribution function of an exponentially distributed random variable with parameter λ.

To simulate a Poisson random variable W with parameter λ, we simply generate a sequence of values of an exponentially distributed random variable with the same parameter, and keep track of the subtotals S_k of these values. We stop generating the sequence when the subtotal first exceeds λ. Assume that we find that

$$S_n \leq \lambda < S_{n+1} \ .$$

Then the value n is returned as a simulated value for W.

Example 5.7 (Queues) Suppose that customers arrive at random times at a service station with one server, and suppose that each customer is served immediately if no one is ahead of him, but must wait his turn in line otherwise. How long should each customer expect to wait? (We define the waiting time of a customer to be the length of time between the time that he arrives and the time that he begins to be served.)

Let us assume that the interarrival times between successive customers are given by random variables $X_1, X_2, \ldots, X_n$ that are mutually independent and identically distributed with an exponential cumulative distribution function given by

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

Let us assume, too, that the service times for successive customers are given by random variables $Y_1, Y_2, \ldots, Y_n$ that again are mutually independent and identically distributed with another exponential cumulative distribution function given by

$$F_Y(t) = 1 - e^{-\mu t}.$$

The parameters λ and μ represent, respectively, the reciprocals of the average time between arrivals of customers and the average service time of the customers. Thus, for example, the larger the value of λ, the smaller the average time between arrivals of customers. We can guess that the length of time a customer will spend in the queue depends on the relative sizes of the average interarrival time and the average service time.

It is easy to verify this conjecture by simulation. The program **Queue** simulates this queueing process. Let $N(t)$ be the number of customers in the queue at time t.

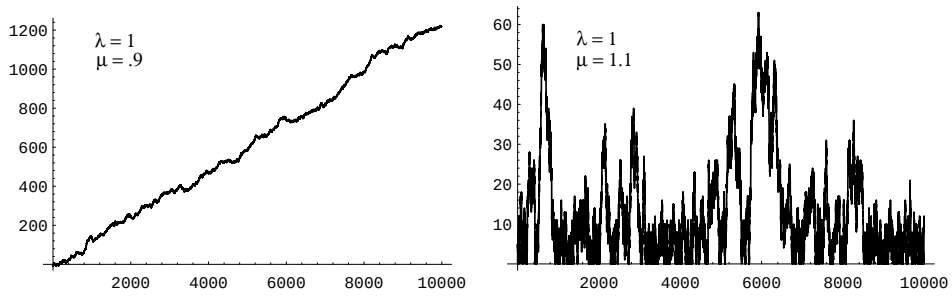

Figure 5.7: Queue sizes.

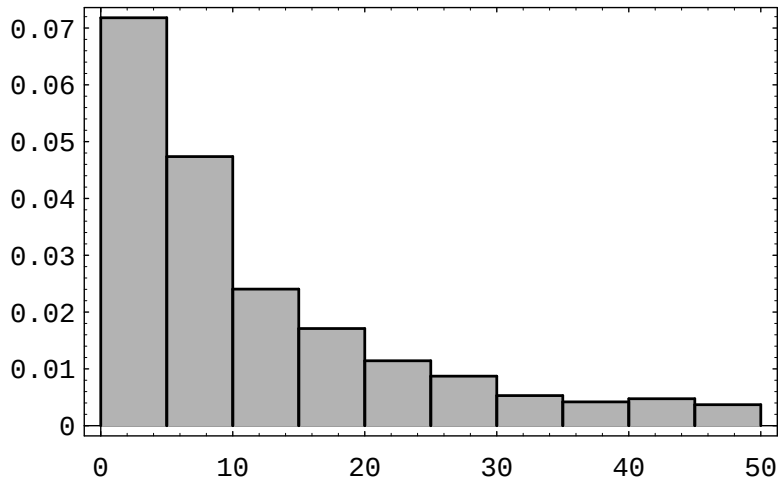

Figure 5.8: Waiting times.

Then we plot $N(t)$ as a function of t for different choices of the parameters λ and μ (see Figure 5.7).

We note that when $\lambda < \mu$, then $1/\lambda > 1/\mu$, so the average interarrival time is greater than the average service time, i.e., customers are served more quickly, on average, than new ones arrive. Thus, in this case, it is reasonable to expect that $N(t)$ remains small. However, if $\lambda > \mu$ then customers arrive more quickly than they are served, and, as expected, $N(t)$ appears to grow without limit.

We can now ask: How long will a customer have to wait in the queue for service? To examine this question, we let W_i be the length of time that the ith customer has to remain in the system (waiting in line and being served). Then we can present these data in a bar graph, using the program **Queue**, to give some idea of how the W_i are distributed (see Figure 5.8). (Here $\lambda = 1$ and $\mu = 1.1$.)

We see that these waiting times appear to be distributed exponentially. This is always the case when $\lambda < \mu$. The proof of this fact is too complicated to give here, but we can verify it by simulation for different choices of λ and μ, as above. □

Functions of a Random Variable

Before continuing our list of important densities, we pause to consider random variables which are functions of other random variables. We will prove a general theorem that will allow us to derive expressions such as Equation 5.5.

Theorem 5.1 Let X be a continuous random variable, and suppose that $\phi(x)$ is a strictly increasing function on the range of X. Define $Y = \phi(X)$. Suppose that X and Y have cumulative distribution functions F_X and F_Y respectively. Then these functions are related by

$$F_Y(y) = F_X(\phi^{-1}(y)).$$

If $\phi(x)$ is strictly decreasing on the range of X, then

$$F_Y(y) = 1 - F_X(\phi^{-1}(y)) \ .$$

Proof. Since ϕ is a strictly increasing function on the range of X, the events $(X \leq \phi^{-1}(y))$ and $(\phi(X) \leq y)$ are equal. Thus, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\phi(X) \leq y) \\ &= P(X \leq \phi^{-1}(y)) \\ &= F_X(\phi^{-1}(y)) \ . \end{aligned}$$

If $\phi(x)$ is strictly decreasing on the range of X, then we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\phi(X) \leq y) \\ &= P(X \geq \phi^{-1}(y)) \\ &= 1 - P(X < \phi^{-1}(y)) \\ &= 1 - F_X(\phi^{-1}(y)) \ . \end{aligned}$$

This completes the proof. □

Corollary 5.1 Let X be a continuous random variable, and suppose that $\phi(x)$ is a strictly increasing function on the range of X. Define $Y = \phi(X)$. Suppose that the density functions of X and Y are f_X and f_Y, respectively. Then these functions are related by

$$f_Y(y) = f_X(\phi^{-1}(y))\frac{d}{dy}\phi^{-1}(y) \ .$$

If $\phi(x)$ is strictly decreasing on the range of X, then

$$f_Y(y) = -f_X(\phi^{-1}(y))\frac{d}{dy}\phi^{-1}(y) \ .$$

Proof. This result follows from Theorem 5.1 by using the Chain Rule. □

If the function ϕ is neither strictly increasing nor strictly decreasing, then the situation is somewhat more complicated but can be treated by the same methods. For example, suppose that $Y = X^2$, Then $\phi(x) = x^2$, and

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(-\sqrt{y} \leq X \leq +\sqrt{y}) \\ &= P(X \leq +\sqrt{y}) - P(X \leq -\sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \ . \end{aligned}$$

Moreover,

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy}(F_X(\sqrt{y}) - F_X(-\sqrt{y})) \\ &= \Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big)\frac{1}{2\sqrt{y}} \ . \end{aligned}$$

We see that in order to express F_Y in terms of F_X when $Y = \phi(X)$, we have to express $P(Y \leq y)$ in terms of $P(X \leq x)$, and this process will depend in general upon the structure of ϕ.

Simulation

Theorem 5.1 tells us, among other things, how to simulate on the computer a random variable Y with a prescribed cumulative distribution function F. We assume that $F(y)$ is strictly increasing for those values of y where $0 < F(y) < 1$. For this purpose, let U be a random variable which is uniformly distributed on $[0, 1]$. Then U has cumulative distribution function $F_U(u) = u$. Now, if F is the prescribed cumulative distribution function for Y, then to write Y in terms of U we first solve the equation

$$F(y) = u$$

for y in terms of u. We obtain $y = F^{-1}(u)$. Note that since F is an increasing function this equation always has a unique solution (see Figure 5.9). Then we set $Z = F^{-1}(U)$ and obtain, by Theorem 5.1,

$$F_Z(y) = F_U(F(y)) = F(y) \ ,$$

since $F_U(u) = u$. Therefore, Z and Y have the same cumulative distribution function. Summarizing, we have the following.

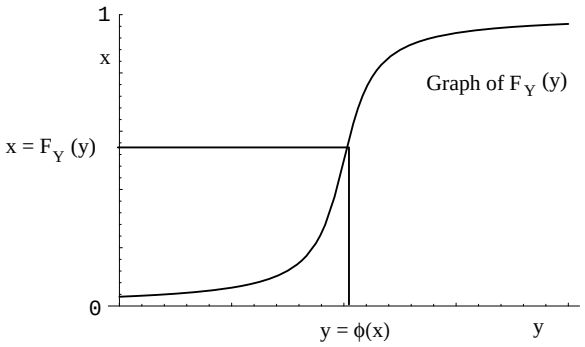

Figure 5.9: Converting a uniform distribution F_U into a prescribed distribution F_Y.

Corollary 5.2 If $F(y)$ is a given cumulative distribution function that is strictly increasing when $0 < F(y) < 1$ and if U is a random variable with uniform distribution on $[0, 1]$, then

$$Y = F^{-1}(U)$$

has the cumulative distribution $F(y)$. □

Thus, to simulate a random variable with a given cumulative distribution F we need only set $Y = F^{-1}(\text{rnd})$.

Normal Density

We now come to the most important density function, the normal density function. We have seen in Chapter 3 that the binomial distribution functions are bell-shaped, even for moderate size values of n. We recall that a binomially-distributed random variable with parameters n and p can be considered to be the sum of n mutually independent 0-1 random variables. A very important theorem in probability theory, called the Central Limit Theorem, states that under very general conditions, if we sum a large number of mutually independent random variables, then the distribution of the sum can be closely approximated by a certain specific continuous density, called the normal density. This theorem will be discussed in Chapter 9.

The normal density function with parameters μ and σ is defined as follows:

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

The parameter μ represents the "center" of the density (and in Chapter 6, we will show that it is the average, or expected, value of the density). The parameter σ is a measure of the "spread" of the density, and thus it is assumed to be positive. (In Chapter 6, we will show that σ is the standard deviation of the density.) We note that it is not at all obvious that the above function is a density, i.e., that its

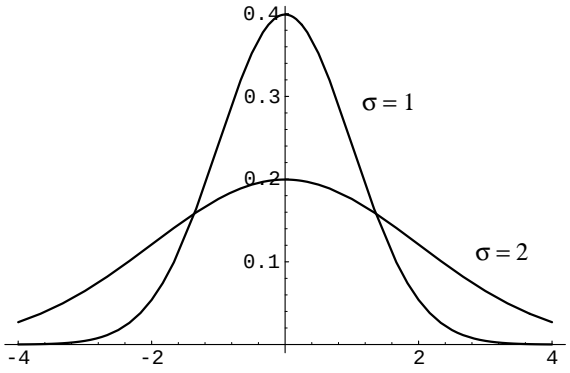

Figure 5.10: Normal density for two sets of parameter values.

integral over the real line equals 1. The cumulative distribution function is given by the formula

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(u-\mu)^2/2\sigma^2}\, du \ .$$

In Figure 5.10 we have included for comparison a plot of the normal density for the cases $\mu = 0$ and $\sigma = 1$, and $\mu = 0$ and $\sigma = 2$.

One cannot write F_X in terms of simple functions. This leads to several problems. First of all, values of F_X must be computed using numerical integration. Extensive tables exist containing values of this function (see Appendix A). Secondly, we cannot write F_X^{-1} in closed form, so we cannot use Corollary 5.2 to help us simulate a normal random variable. For this reason, special methods have been developed for simulating a normal distribution. One such method relies on the fact that if U and V are independent random variables with uniform densities on $[0, 1]$, then the random variables

$$X = \sqrt{-2\log U}\cos 2\pi V$$

and

$$Y = \sqrt{-2\log U}\sin 2\pi V$$

are independent, and have normal density functions with parameters $\mu = 0$ and $\sigma = 1$. (This is not obvious, nor shall we prove it here. See Box and Muller.[9])

Let Z be a normal random variable with parameters $\mu = 0$ and $\sigma = 1$. A normal random variable with these parameters is said to be a *standard* normal random variable. It is an important and useful fact that if we write

$$X = \sigma Z + \mu \ ,$$

then X is a normal random variable with parameters μ and σ. To show this, we will use Theorem 5.1. We have $\phi(z) = \sigma z + \mu$, $\phi^{-1}(x) = (x - \mu)/\sigma$, and

$$F_X(x) \quad = \quad F_Z\left(\frac{x-\mu}{\sigma}\right),$$

[9]G. E. P. Box and M. E. Muller, *A Note on the Generation of Random Normal Deviates*, Ann. of Math. Stat. 29 (1958), pgs. 610-611.

$$\begin{aligned} f_X(x) &= f_Z\left(\frac{x-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \ . \end{aligned}$$

The reader will note that this last expression is the density function with parameters μ and σ, as claimed.

We have seen above that it is possible to simulate a standard normal random variable Z. If we wish to simulate a normal random variable X with parameters μ and σ, then we need only transform the simulated values for Z using the equation $X = \sigma Z + \mu$.

Suppose that we wish to calculate the value of a cumulative distribution function for the normal random variable X, with parameters μ and σ. We can reduce this calculation to one concerning the standard normal random variable Z as follows:

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \ . \end{aligned}$$

This last expression can be found in a table of values of the cumulative distribution function for a standard normal random variable. Thus, we see that it is unnecessary to make tables of normal distribution functions with arbitrary μ and σ.

The process of changing a normal random variable to a standard normal random variable is known as standardization. If X has a normal distribution with parameters μ and σ and if

$$Z = \frac{X-\mu}{\sigma} \ ,$$

then Z is said to be the standardized version of X.

The following example shows how we use the standardized version of a normal random variable X to compute specific probabilities relating to X.

Example 5.8 Suppose that X is a normally distributed random variable with parameters $\mu = 10$ and $\sigma = 3$. Find the probability that X is between 4 and 16.

To solve this problem, we note that $Z = (X - 10)/3$ is the standardized version of X. So, we have

$$\begin{aligned} P(4 \leq X \leq 16) &= P(X \leq 16) - P(X \leq 4) \\ &= F_X(16) - F_X(4) \\ &= F_Z\left(\frac{16-10}{3}\right) - F_Z\left(\frac{4-10}{3}\right) \\ &= F_Z(2) - F_Z(-2) \ . \end{aligned}$$

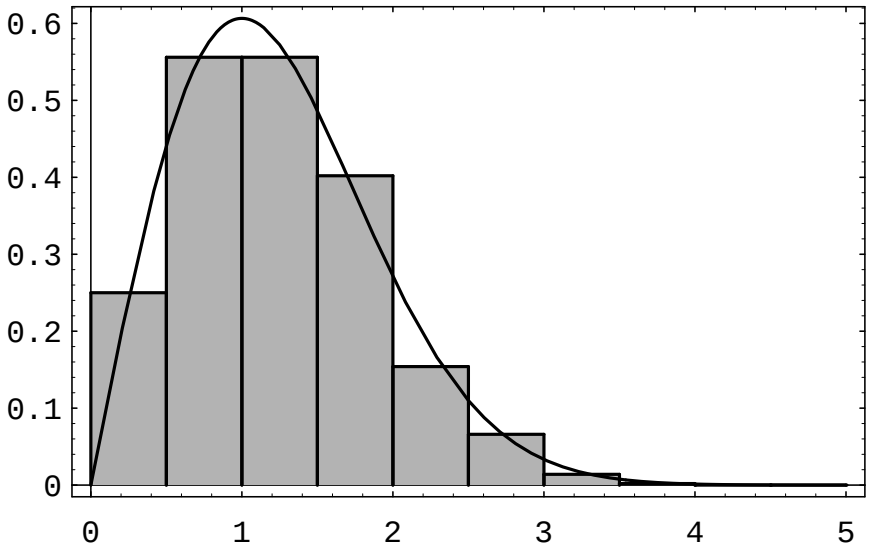

Figure 5.11: Distribution of dart distances in 1000 drops.

This last expression can be evaluated by using tabulated values of the standard normal distribution function (see 11.5); when we use this table, we find that $F_Z(2) = .9772$ and $F_Z(-2) = .0228$. Thus, the answer is .9544.

In Chapter 6, we will see that the parameter μ is the mean, or average value, of the random variable X. The parameter σ is a measure of the spread of the random variable, and is called the standard deviation. Thus, the question asked in this example is of a typical type, namely, what is the probability that a random variable has a value within two standard deviations of its average value. □

Maxwell and Rayleigh Densities

Example 5.9 Suppose that we drop a dart on a large table top, which we consider as the xy-plane, and suppose that the x and y coordinates of the dart point are independent and have a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. How is the distance of the point from the origin distributed?

This problem arises in physics when it is assumed that a moving particle in R^n has components of the velocity that are mutually independent and normally distributed and it is desired to find the density of the speed of the particle. The density in the case $n = 3$ is called the Maxwell density.

The density in the case $n = 2$ (i.e. the dart board experiment described above) is called the Rayleigh density. We can simulate this case by picking independently a pair of coordinates (x, y), each from a normal distribution with $\mu = 0$ and $\sigma = 1$ on $(-\infty, \infty)$, calculating the distance $r = \sqrt{x^2 + y^2}$ of the point (x, y) from the origin, repeating this process a large number of times, and then presenting the results in a bar graph. The results are shown in Figure 5.11.

	Female	Male	
A	37	56	93
B	63	60	123
C	47	43	90
Below C	5	8	13
	152	167	319

Table 5.8: Calculus class data.

	Female	Male	
A	44.3	48.7	93
B	58.6	64.4	123
C	42.9	47.1	90
Below C	6.2	6.8	13
	152	167	319

Table 5.9: Expected data.

We have also plotted the theoretical density

$$f(r) = re^{-r^2/2} \ .$$

This will be derived in Chapter 7; see Example 7.7. □

Chi-Squared Density

We return to the problem of independence of traits discussed in Example 5.6. It is frequently the case that we have two traits, each of which have several different values. As was seen in the example, quite a lot of calculation was needed even in the case of two values for each trait. We now give another method for testing independence of traits, which involves much less calculation.

Example 5.10 Suppose that we have the data shown in Table 5.8 concerning grades and gender of students in a Calculus class. We can use the same sort of model in this situation as was used in Example 5.6. We imagine that we have an urn with 319 balls of two colors, say blue and red, corresponding to females and males, respectively. We now draw 93 balls, without replacement, from the urn. These balls correspond to the grade of A. We continue by drawing 123 balls, which correspond to the grade of B. When we finish, we have four sets of balls, with each ball belonging to exactly one set. (We could have stipulated that the balls were of four colors, corresponding to the four possible grades. In this case, we would draw a subset of size 152, which would correspond to the females. The balls remaining in the urn would correspond to the males. The choice does not affect the final determination of whether we should reject the hypothesis of independence of traits.)

The expected data set can be determined in exactly the same way as in Example 5.6. If we do this, we obtain the expected values shown in Table 5.9. Even if

the traits are independent, we would still expect to see some differences between the numbers in corresponding boxes in the two tables. However, if the differences are large, then we might suspect that the two traits are not independent. In Example 5.6, we used the probability distribution of the various possible data sets to compute the probability of finding a data set that differs from the expected data set by at least as much as the actual data set does. We could do the same in this case, but the amount of computation is enormous.

Instead, we will describe a single number which does a good job of measuring how far a given data set is from the expected one. To quantify how far apart the two sets of numbers are, we could sum the squares of the differences of the corresponding numbers. (We could also sum the absolute values of the differences, but we would not want to sum the differences.) Suppose that we have data in which we expect to see 10 objects of a certain type, but instead we see 18, while in another case we expect to see 50 objects of a certain type, but instead we see 58. Even though the two differences are about the same, the first difference is more surprising than the second, since the expected number of outcomes in the second case is quite a bit larger than the expected number in the first case. One way to correct for this is to divide the individual squares of the differences by the expected number for that box. Thus, if we label the values in the eight boxes in the first table by O_i (for observed values) and the values in the eight boxes in the second table by E_i (for expected values), then the following expression might be a reasonable one to use to measure how far the observed data is from what is expected:

$$\sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} \ .$$

This expression is a random variable, which is usually denoted by the symbol χ^2, pronounced "ki-squared." It is called this because, under the assumption of independence of the two traits, the density of this random variable can be computed and is approximately equal to a density called the chi-squared density. We choose not to give the explicit expression for this density, since it involves the gamma function, which we have not discussed. The chi-squared density is, in fact, a special case of the general gamma density.

In applying the chi-squared density, tables of values of this density are used, as in the case of the normal density. The chi-squared density has one parameter n, which is called the number of degrees of freedom. The number n is usually easy to determine from the problem at hand. For example, if we are checking two traits for independence, and the two traits have a and b values, respectively, then the number of degrees of freedom of the random variable χ^2 is $(a-1)(b-1)$. So, in the example at hand, the number of degrees of freedom is 3.

We recall that in this example, we are trying to test for independence of the two traits of gender and grades. If we assume these traits are independent, then the ball-and-urn model given above gives us a way to simulate the experiment. Using a computer, we have performed 1000 experiments, and for each one, we have calculated a value of the random variable χ^2. The results are shown in Figure 5.12, together with the chi-squared density function with three degrees of freedom.

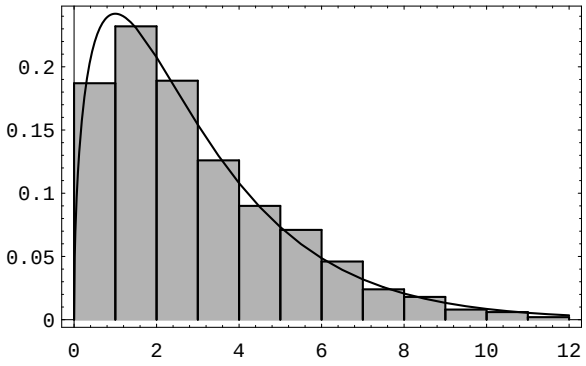

Figure 5.12: Chi-squared density with three degrees of freedom.

As we stated above, if the value of the random variable χ^2 is large, then we would tend not to believe that the two traits are independent. But how large is large? The actual value of this random variable for the data above is 4.13. In Figure 5.12, we have shown the chi-squared density with 3 degrees of freedom. It can be seen that the value 4.13 is larger than most of the values taken on by this random variable.

Typically, a statistician will compute the value v of the random variable χ^2, just as we have done. Then, by looking in a table of values of the chi-squared density, a value v_0 is determined which is only exceeded 5% of the time. If $v \geq v_0$, the statistician rejects the hypothesis that the two traits are independent. In the present case, $v_0 = 7.815$, so we would not reject the hypothesis that the two traits are independent. □

Cauchy Density

The following example is from Feller.[10]

Example 5.11 Suppose that a mirror is mounted on a vertical axis, and is free to revolve about that axis. The axis of the mirror is 1 foot from a straight wall of infinite length. A pulse of light is shown onto the mirror, and the reflected ray hits the wall. Let ϕ be the angle between the reflected ray and the line that is perpendicular to the wall and that runs through the axis of the mirror. We assume that ϕ is uniformly distributed between $-\pi/2$ and $\pi/2$. Let X represent the distance between the point on the wall that is hit by the reflected ray and the point on the wall that is closest to the axis of the mirror. We now determine the density of X.

Let B be a fixed positive quantity. Then $X \geq B$ if and only if $\tan(\phi) \geq B$, which happens if and only if $\phi \geq \arctan(B)$. This happens with probability

$$\frac{\pi/2 - \arctan(B)}{\pi} .$$

[10]W. Feller, *An Introduction to Probability Theory and Its Applications,*, vol. 2, (New York: Wiley, 1966)

Thus, for positive B, the cumulative distribution function of X is

$$F(B) = 1 - \frac{\pi/2 - \arctan(B)}{\pi} \ .$$

Therefore, the density function for positive B is

$$f(B) = \frac{1}{\pi(1+B^2)} \ .$$

Since the physical situation is symmetric with respect to $\phi = 0$, it is easy to see that the above expression for the density is correct for negative values of B as well.

The Law of Large Numbers, which we will discuss in Chapter 8, states that in many cases, if we take the average of independent values of a random variable, then the average approaches a specific number as the number of values increases. It turns out that if one does this with a Cauchy-distributed random variable, the average does not approach any specific number. □

Exercises

1 Choose a number U from the unit interval $[0, 1]$ with uniform distribution. Find the cumulative distribution and density for the random variables

(a) $Y = U + 2$.

(b) $Y = U^3$.

2 Choose a number U from the interval $[0, 1]$ with uniform distribution. Find the cumulative distribution and density for the random variables

(a) $Y = 1/(U + 1)$.

(b) $Y = \log(U + 1)$.

3 Use Corollary 5.2 to derive the expression for the random variable given in Equation 5.5. *Hint*: The random variables $1 - rnd$ and rnd are identically distributed.

4 Suppose we know a random variable Y as a function of the uniform random variable U: $Y = \phi(U)$, and suppose we have calculated the cumulative distribution function $F_Y(y)$ and thence the density $f_Y(y)$. How can we check whether our answer is correct? An easy simulation provides the answer: Make a bar graph of $Y = \phi(rnd)$ and compare the result with the graph of $f_Y(y)$. These graphs should look similar. Check your answers to Exercises 1 and 2 by this method.

5 Choose a number U from the interval $[0, 1]$ with uniform distribution. Find the cumulative distribution and density for the random variables

(a) $Y = |U - 1/2|$.

(b) $Y = (U - 1/2)^2$.

6 Check your results for Exercise 5 by simulation as described in Exercise 4.

7 Explain how you can generate a random variable whose cumulative distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ x^2, & \text{if } 0 \leq x \leq 1, \\ 1, & \text{if } x > 1. \end{cases}$$

8 Write a program to generate a sample of 1000 random outcomes each of which is chosen from the distribution given in Exercise 7. Plot a bar graph of your results and compare this empirical density with the density for the cumulative distribution given in Exercise 7.

9 Let U, V be random numbers chosen independently from the interval $[0, 1]$ with uniform distribution. Find the cumulative distribution and density of each of the variables

(a) $Y = U + V$.

(b) $Y = |U - V|$.

10 Let U, V be random numbers chosen independently from the interval $[0, 1]$. Find the cumulative distribution and density for the random variables

(a) $Y = \max(U, V)$.

(b) $Y = \min(U, V)$.

11 Write a program to simulate the random variables of Exercises 9 and 10 and plot a bar graph of the results. Compare the resulting empirical density with the density found in Exercises 9 and 10.

12 A number U is chosen at random in the interval $[0, 1]$. Find the probability that

(a) $R = U^2 < 1/4$.

(b) $S = U(1 - U) < 1/4$.

(c) $T = U/(1 - U) < 1/4$.

13 Find the cumulative distribution function F and the density function f for each of the random variables R, S, and T in Exercise 12.

14 A point P in the unit square has coordinates X and Y chosen at random in the interval $[0, 1]$. Let D be the distance from P to the nearest edge of the square, and E the distance to the nearest corner. What is the probability that

(a) $D < 1/4$?

(b) $E < 1/4$?

15 In Exercise 14 find the cumulative distribution F and density f for the random variable D.

16 Let X be a random variable with density function

$$f_X(x) = \begin{cases} cx(1-x), & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of c?

(b) What is the cumulative distribution function F_X for X?

(c) What is the probability that $X < 1/4$?

17 Let X be a random variable with cumulative distribution function

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ \sin^2(\pi x/2), & \text{if } 0 \le x \le 1, \\ 1, & \text{if } 1 < x. \end{cases}$$

(a) What is the density function f_X for X?

(b) What is the probability that $X < 1/4$?

18 Let X be a random variable with cumulative distribution function F_X, and let $Y = X + b$, $Z = aX$, and $W = aX + b$, where a and b are any constants. Find the cumulative distribution functions F_Y, F_Z, and F_W. *Hint*: The cases $a > 0$, $a = 0$, and $a < 0$ require different arguments.

19 Let X be a random variable with density function f_X, and let $Y = X + b$, $Z = aX$, and $W = aX + b$, where $a \neq 0$. Find the density functions f_Y, f_Z, and f_W. (See Exercise 18.)

20 Let X be a random variable uniformly distributed over $[c, d]$, and let $Y = aX + b$. For what choice of a and b is Y uniformly distributed over $[0, 1]$?

21 Let X be a random variable with cumulative distribution function F strictly increasing on the range of X. Let $Y = F(X)$. Show that Y is uniformly distributed in the interval $[0, 1]$. (The formula $X = F^{-1}(Y)$ then tells us how to construct X from a uniform random variable Y.)

22 Let X be a random variable with cumulative distribution function F. The *median* of X is the value m for which $F(m) = 1/2$. Then $X < m$ with probability $1/2$ and $X > m$ with probability $1/2$. Find m if X is

(a) uniformly distributed over the interval $[a, b]$.

(b) normally distributed with parameters μ and σ.

(c) exponentially distributed with parameter λ.

23 Let X be a random variable with density function f_X. The *mean* of X is the value $\mu = \int x f_x(x)\, dx$. Then μ gives an average value for X (see Section 6.3). Find μ if X is distributed uniformly, normally, or exponentially, as in Exercise 22.

Test Score	Letter grade
$\mu + \sigma < x$	A
$\mu < x < \mu + \sigma$	B
$\mu - \sigma < x < \mu$	C
$\mu - 2\sigma < x < \mu - \sigma$	D
$x < \mu - 2\sigma$	F

Table 5.10: Grading on the curve.

24 Let X be a random variable with density function f_X. The *mode* of X is the value M for which $f(M)$ is maximum. Then values of X near M are most likely to occur. Find M if X is distributed normally or exponentially, as in Exercise 22. What happens if X is distributed uniformly?

25 Let X be a random variable normally distributed with parameters $\mu = 70$, $\sigma = 10$. Estimate

(a) $P(X > 50)$.

(b) $P(X < 60)$.

(c) $P(X > 90)$.

(d) $P(60 < X < 80)$.

26 Bridies' Bearing Works manufactures bearing shafts whose diameters are normally distributed with parameters $\mu = 1$, $\sigma = .002$. The buyer's specifications require these diameters to be $1.000 \pm .003$ cm. What fraction of the manufacturer's shafts are likely to be rejected? If the manufacturer improves her quality control, she can reduce the value of σ. What value of σ will ensure that no more than 1 percent of her shafts are likely to be rejected?

27 A final examination at Podunk University is constructed so that the test scores are approximately normally distributed, with parameters μ and σ. The instructor assigns letter grades to the test scores as shown in Table 5.10 (this is the process of "grading on the curve").

What fraction of the class gets A, B, C, D, F?

28 (Ross[11]) An expert witness in a paternity suit testifies that the length (in days) of a pregnancy, from conception to delivery, is approximately normally distributed, with parameters $\mu = 270$, $\sigma = 10$. The defendant in the suit is able to prove that he was out of the country during the period from 290 to 240 days before the birth of the child. What is the probability that the defendant was in the country when the child was conceived?

29 Suppose that the time (in hours) required to repair a car is an exponentially distributed random variable with parameter $\lambda = 1/2$. What is the probability that the repair time exceeds 4 hours? If it exceeds 4 hours what is the probability that it exceeds 8 hours?

[11]S. Ross, *A First Course in Probability Theory,* 2d ed. (New York: Macmillan, 1984).

30 Suppose that the number of years a car will run is exponentially distributed with parameter $\mu = 1/4$. If Prosser buys a used car today, what is the probability that it will still run after 4 years?

31 Let U be a uniformly distributed random variable on $[0, 1]$. What is the probability that the equation

$$x^2 + 4Ux + 1 = 0$$

has two distinct real roots x_1 and x_2?

32 Write a program to simulate the random variables whose densities are given by the following, making a suitable bar graph of each and comparing the exact density with the bar graph.

(a) $f_X(x) = e^{-x}$ on $[0, \infty)$ (but just do it on $[0, 10]$).

(b) $f_X(x) = 2x$ on $[0, 1]$.

(c) $f_X(x) = 3x^2$ on $[0, 1]$.

(d) $f_X(x) = 4|x - 1/2|$ on $[0, 1]$.

33 Suppose we are observing a process such that the time between occurrences is exponentially distributed with $\lambda = 1/30$ (i.e., the average time between occurrences is 30 minutes). Suppose that the process starts at a certain time and we start observing the process 3 hours later. Write a program to simulate this process. Let T denote the length of time that we have to wait, after we start our observation, for an occurrence. Have your program keep track of T. What is an estimate for the average value of T?

34 Jones puts in two new lightbulbs: a 60 watt bulb and a 100 watt bulb. It is claimed that the lifetime of the 60 watt bulb has an exponential density with average lifetime 200 hours ($\lambda = 1/200$). The 100 watt bulb also has an exponential density but with average lifetime of only 100 hours ($\lambda = 1/100$). Jones wonders what is the probability that the 100 watt bulb will outlast the 60 watt bulb.

If X and Y are two independent random variables with exponential densities $f(x) = \lambda e^{-\lambda x}$ and $g(x) = \mu e^{-\mu x}$, respectively, then the probability that X is less than Y is given by

$$P(X < Y) = \int_0^\infty f(x)(1 - G(x))\, dx,$$

where $G(x)$ is the cumulative distribution function for $g(x)$. Explain why this is the case. Use this to show that

$$P(X < Y) = \frac{\lambda}{\lambda + \mu}$$

and to answer Jones's question.

35 Consider the simple queueing process of Example 5.7. Suppose that you watch the size of the queue. If there are j people in the queue the next time the queue size changes it will either decrease to $j-1$ or increase to $j+1$. Use the result of Exercise 34 to show that the probability that the queue size decreases to $j-1$ is $\mu/(\mu+\lambda)$ and the probability that it increases to $j+1$ is $\lambda/(\mu+\lambda)$. When the queue size is 0 it can only increase to 1. Write a program to simulate the queue size. Use this simulation to help formulate a conjecture containing conditions on μ and λ that will ensure that the queue will have times when it is empty.

36 Let X be a random variable having an exponential density with parameter λ. Find the density for the random variable $Y = rX$, where r is a positive real number.

37 Let X be a random variable having a normal density and consider the random variable $Y = e^X$. Then Y has a *log normal* density. Find this density of Y.

38 Let X_1 and X_2 be independent random variables and for $i = 1, 2$, let $Y_i = \phi_i(X_i)$, where ϕ_i is strictly increasing on the range of X_i. Show that Y_1 and Y_2 are independent. Note that the same result is true without the assumption that the ϕ_i's are strictly increasing, but the proof is more difficult.