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

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

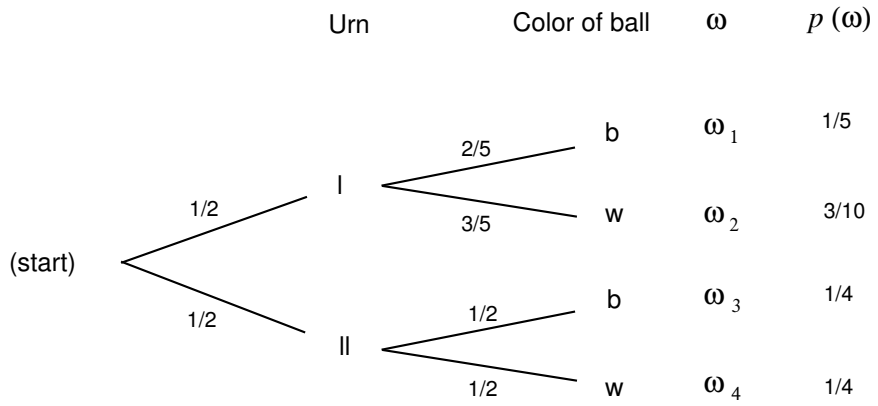

Figure 4.1: Tree diagram.

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

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

□

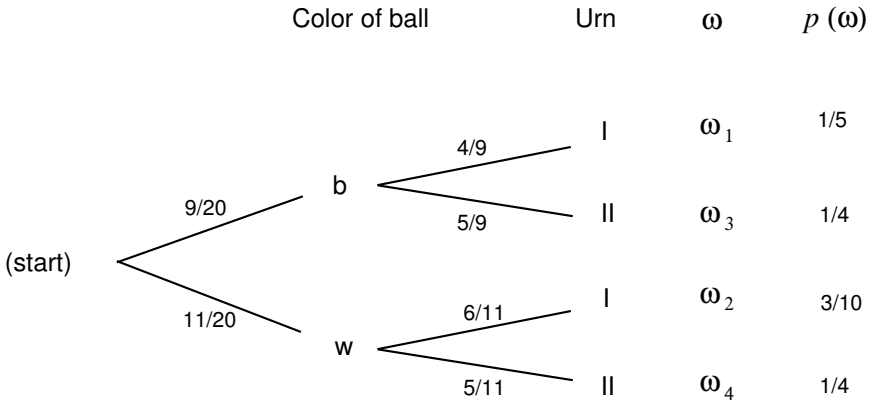

Figure 4.2: Reverse tree diagram.

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

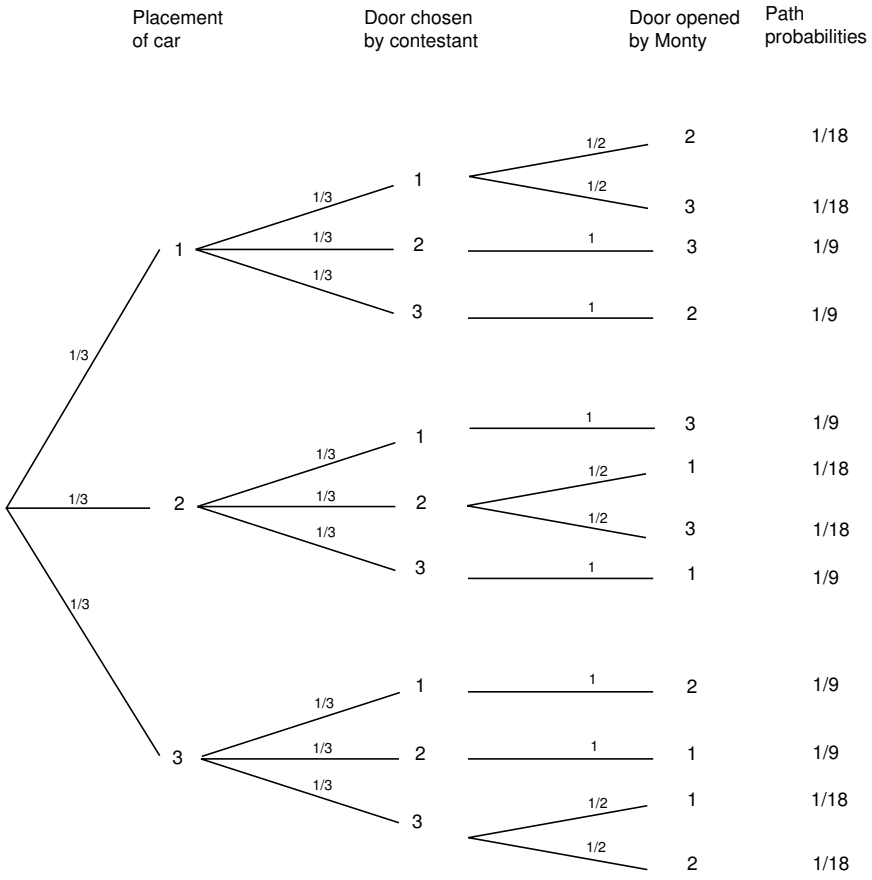

Figure 4.3: The Monty Hall problem.

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

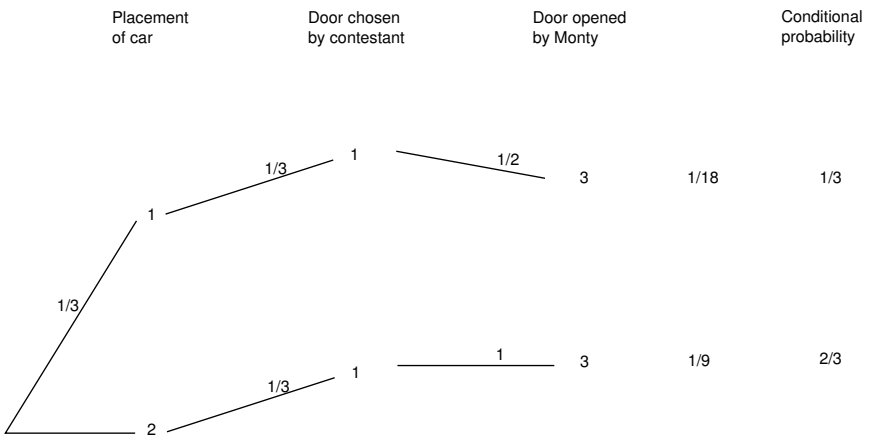

Figure 4.4: Conditional probabilities for the Monty Hall problem.

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, $\ldots$, 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 $\bar{A}_1$, $\bar{A}_2$, $\ldots$, $\bar{A}_n$ 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, $\ldots$, 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: