# Chapter 4

# Conditional Probability

## 4.1 Introduction

In the previous chapter we determined the probabilities for some simple experiments. An example was the die toss that produced a number from 1 to 6 "at random". Hence, a probability of 1/6 was assigned to each possible outcome. In many real-world "experiments", the outcomes are not completely random since we have some prior knowledge. For instance, knowing that it has rained the previous 2 days might influence our assignment of the probability of sunshine for the following day. Another example is to determine the probability that an individual chosen from some general population weighs more than 200 lbs., knowing that his height exceeds 6 ft. This motivates our interest in how to determine the probability of an event, given that we have some prior knowledge. For the die tossing experiment we might inquire as to the probability of obtaining a 4, if it is known that the outcome is an even number. The additional knowledge should undoubtedly change our probability assignments. For example, if it is known that the outcome is an even number, then the probability of any odd-numbered outcome must be zero. It is this interaction between the original probabilities and the probabilities in light of prior knowledge that we wish to describe and quantify, leading to the concept of a *conditional probability.*

## 4.2 Summary

Section 4.3 motivates and then defines the conditional probability as (4.1). In doing so the concept of a joint event and its probability are introduced as well as the marginal probability of (4.3). Conditional probabilities can be greater than, less than, or equal to the ordinary probability as illustrated in Figure 4.2. Also, conditional probabilities are true probabilities in that they satisfy the basic axioms and so can be manipulated in the usual ways. Using the law of total probability (4.4), the probabilities for compound experiments are easily determined. When the conditional probability is equal to the ordinary probability, the events are said to be statistically independent. Then, knowledge of the occurrence of one event does not change the probability of the other event. The condition for two events to be independent is given by (4.5). Three events are statistically independent if the conditions (4.6)–(4.9) hold. Bayes' theorem is defined by either (4.13) or (4.14). Embodied in the theorem are the concepts of a prior probability (before the experiment is conducted) and a posterior probability (after the experiment is conducted). Conclusions may be drawn based on the outcome of an experiment as to whether certain hypotheses are true. When an experiment is repeated multiple times and the experiments are independent, the probability of a joint event is easily found via (4.15). Some probability laws that result from the independent multiple experiment assumption are the binomial (4.16), the geometric (4.17), and the multinomial (4.19). For dependent multiple experiments (4.20) must be used to determine probabilities of joint events. If, however, the experimental outcomes probabilities only depend on the previous experimental outcome, then the Markov condition is satisfied. This results in the simpler formula for determining joint probabilities given by (4.21). Also, this assumption leads to the concept of a Markov chain, an example of which is shown in Figure 4.8. Finally, in Section 4.7 an example of the use of Bayes' theorem to detect the presence of a cluster is investigated.

## 4.3 Joint Events and the Conditional Probability

In formulating a useful theory of conditional probability we are led to consider two events. Event $A$ is our event of interest while event $B$ represents the event that embodies our prior knowledge. For the fair die toss example described in the introduction, the event of interest is $A = \{4\}$ and the event describing our prior knowledge is an even outcome or $B = \{2,4,6\}$. Note that when we say that the outcome must be even, we do not elaborate on why this is the case. It may be because someone has observed the outcome of the experiment and conveyed this partial information to us. Alternatively, it may be that the experimenter loathes odd outcomes, and therefore keeps tossing the die until an even outcome is obtained. Conditional probability does not address the reasons for the prior information, only how to accommodate it into a probabilistic framework. Continuing with the fair die example, a typical sequence of outcomes for a repeated experiment is shown in Figure 4.1. The odd outcomes are shown as dashed lines and are to be ignored. From the figure we see that the probability of a 4 is about $9/25 = 0.36$, or about $1/3$, using a relative frequency interpretation of probability. This has been found by taking the total number of 4's and dividing by the total number of 2's, 4's, and 6's. Specifically, we have that

$$\frac{N_A}{N_B} = \frac{9}{25}.$$

Another problem might be to determine the probability of $A = \{1,4\}$, knowing that the outcome is even. In this case, we should use $N_{A\cap B}/N_B$ to make sure we

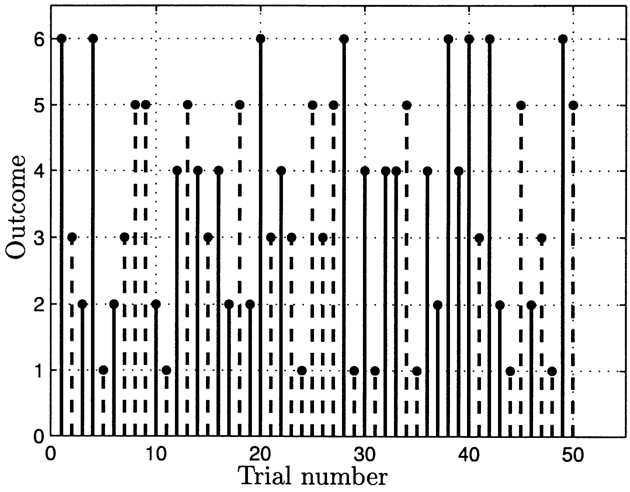


Figure 4.1: Outcomes for repeated tossing of a fair die.

only count the outcomes that can occur in light of our knowledge of $B$. For this example, only the 4 in $\{1,4\}$ could have occurred. If an outcome is not in $B$, then that outcome will not be included in $A \cap B$ and will not be counted in $N_{A\cap B}$. Now letting $\mathcal{S} = \{1,2,3,4,5,6\}$ be the sample space and $N_{\mathcal{S}}$ its size, the probability of $A$ given $B$ is

$$\frac{N_{A\cap B}}{N_B} = \frac{\frac{N_{A\cap B}}{N_{\mathcal{S}}}}{\frac{N_B}{N_{\mathcal{S}}}} \approx \frac{P[A\cap B]}{P[B]}.$$

This is termed the *conditional probability* and is denoted by $P[A|B]$ so that we have as our definition

$$P[A|B] = \frac{P[A\cap B]}{P[B]}. \tag{4.1}$$

Note that to determine it, we require $P[A\cap B]$ which is the probability of both $A$ and $B$ occurring or the probability of the intersection. Intuitively, the conditional probability is the proportion of time $A$ and $B$ occurs divided by the proportion of time that $B$ occurs. The event $B = \{2,4,6\}$ comprises a new sample space and is sometimes called the *reduced sample space.* The denominator term in (4.1) serves to normalize the conditional probabilities so that the probability of the reduced sample space is one (set $A = B$ in (4.1)). Returning to the die toss, the probability of a 4, given that the outcome is even, is found as

$$\begin{aligned} A\cap B &= \{4\}\cap\{2,4,6\} = \{4\} = A \\ B &= \{2,4,6\} \end{aligned}$$

| | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ | $P[H_i]$ |
|---|---|---|---|---|---|---|
| | 100–130 | 130–160 | 160–190 | 190–220 | 220–250 | |
| $H_1$ $5'-5'4''$ | 0.08 | 0.04 | 0.02 | 0 | 0 | 0.14 |
| $H_2$ $5'4''-5'8''$ | 0.06 | 0.12 | 0.06 | 0.02 | 0 | 0.26 |
| $H_3$ $5'8''-6'$ | 0 | 0.06 | 0.14 | 0.06 | 0 | 0.26 |
| $H_4$ $6'-6'4''$ | 0 | 0.02 | 0.06 | 0.10 | 0.04 | 0.22 |
| $H_5$ $6'4''-6'8''$ | 0 | 0 | 0 | 0.08 | 0.04 | 0.12 |

Table 4.1: Joint probabilities for heights and weights of college students.

and therefore

$$\begin{aligned} P[A|B] &= \frac{P[A \cap B]}{P[B]} = \frac{P[A]}{P[B]} \\ &= \frac{1/6}{3/6} = \frac{1}{3} \end{aligned}$$

as expected. Note that $P[A \cap B]$ and $P[B]$ are computed based on the *original sample space*, $\mathcal{S}$.

The event $A \cap B$ is usually called the *joint event* since both events must occur for a nonempty intersection. Likewise, $P[A \cap B]$ is termed the *joint probability*, but of course, it is nothing more than the probability of an intersection. Also, $P[A]$ is called the *marginal probability* to distinguish it from the joint and conditional probabilities. The reason for this terminology will be discussed shortly.

In defining the conditional probability of (4.1) it is assumed that $P[B] \neq 0$. Otherwise, theoretically and practically, the definition would not make sense. Another example follows.

### Example 4.1 – Heights and weights of college students

A population of college students have heights $H$ and weights $W$ which are grouped into ranges as shown in Table 4.1. The table gives the joint probability of a student having a given height and weight, which can be denoted as $P[H_i \cap W_j]$. For example, if a student is selected, the probability of his/her height being between $5'4''$ and $5'8''$ and also his/her weight being between 130 lbs. and 160 lbs. is 0.12. Now consider the event that the student has a weight in the range 130–160 lbs. Calling this event $A$ we next determine its probability. Since $A = \{(H, W) : H = H_1, \ldots, H_5; W = W_2\}$, it is explicitly

$$A = \{(H_1, W_2), (H_2, W_2), (H_3, W_2), (H_4, W_2), (H_5, W_2)\}$$

and since the simple events are by definition mutually exclusive, we have by Axiom

3′ (see Section 3.4)

$$\begin{aligned} P[A] &= \sum_{i=1}^{5} P[(H_i, W_2)] = 0.04 + 0.12 + 0.06 + 0.02 + 0 \\ &= 0.24. \end{aligned}$$

Next we determine the probability that a student's weight is in the range of 130–160 lbs., *given* that the student has height less than $6'$. The event of interest $A$ is the same as before. The conditioning event is $B = \{(H, W) : H = H_1, H_2, H_3; W = W_1, \ldots, W_5\}$ so that $A \cap B = \{(H_1, W_2), (H_2, W_2), (H_3, W_2)\}$ and

$$\begin{aligned} P[A|B] &= \frac{P[A \cap B]}{P[B]} = \frac{0.04 + 0.12 + 0.06}{0.14 + 0.26 + 0.26} \\ &= 0.33. \end{aligned}$$

We see that it is more probable that the student has weight between 130 and 160 lbs. if it is known beforehand that his/her height is less than $6'$. Note that in finding $P[B]$ we have used

$$P[B] = \sum_{i=1}^{3} \sum_{j=1}^{5} P[(H_i, W_j)] \tag{4.2}$$

which is determined by first summing along each row to produce the entries shown in Table 4.1 as $P[H_i]$. These are given by

$$P[H_i] = \sum_{j=1}^{5} P[(H_i, W_j)] \tag{4.3}$$

and then summing the $P[H_i]$'s for $i = 1, 2, 3$. Hence, we could have written (4.2) equivalently as

$$P[B] = \sum_{i=1}^{3} P[H_i].$$

The probabilities $P[H_i]$ are called the *marginal probabilities* since they are written in the *margin* of the table. If we were to sum along the columns, then we would obtain the marginal probabilities for the weights or $P[W_j]$. These are given by

$$P[W_j] = \sum_{i=1}^{5} P[(H_i, W_j)].$$

It is important to observe that by utilizing the information that the student's height is less than $6'$, the probability of the event has changed; in this case, it has increased from 0.24 to 0.33. It is also possible that the opposite may occur. If we were to determine the probability that the student's weight is in the range

130–160 lbs., given that he/she has a height *greater* than 6$'$, then defining the conditioning event as $B = \{(H, W) : H = H_4, H_5; W = W_1, \ldots, W_5\}$ and noting that $A \cap B = \{(H_4, W_2), (H_5, W_2\}$ we have

$$\begin{aligned} P[A|B] &= \frac{0.02 + 0}{0.22 + 0.12} \\ &= 0.058. \end{aligned}$$

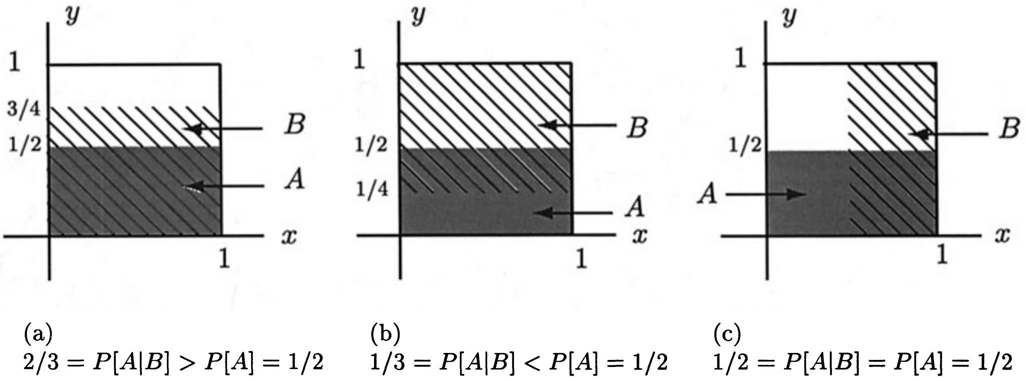
Hence, the conditional probability has now decreased with respect to the unconditional probability or $P[A]$.

◇

In general we may have

$$\begin{aligned} P[A|B] &> P[A] \\ P[A|B] &< P[A] \\ P[A|B] &= P[A]. \end{aligned}$$

See Figure 4.2 for another example. The last possibility is of particular interest since it states that the probability of an event $A$ is the same whether or not we know that $B$ has occurred. In this case, the event $A$ is said to be *statistically independent* of the event $B$. In the next section, we will explore this further.



(a)
$2/3 = P[A|B] > P[A] = 1/2$

(b)
$1/3 = P[A|B] < P[A] = 1/2$

(c)
$1/2 = P[A|B] = P[A] = 1/2$

Figure 4.2: Illustration of possible relationships of conditional probability to ordinary probability.

Before proceeding, we wish to emphasize that a conditional probability is a true probability in that it satisfies the axioms described in Chapter 3. As a result, all the rules that allow one to manipulate probabilities also apply to conditional probabilities. For example, since Property 3.1 must hold, it follows that $P[A^c|B] = 1-P[A|B]$ (see also Problem 4.10). To prove that the axioms are satisfied for conditional probabilities we first assume that the axioms hold for ordinary probabilities. Then,

**Axiom 1**

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \geq 0$$

since $P[A \cap B] \geq 0$ and $P[B] \geq 0$.

**Axiom 2**

$$P[\mathcal{S}|B] = \frac{P[\mathcal{S} \cap B]}{P[B]} = \frac{P[B]}{P[B]} = 1.$$

**Axiom 3** If $A$ and $C$ are mutually exclusive events, then

$$\begin{aligned}
P[A \cup C|B] &= \frac{P[(A \cup C) \cap B]}{P[B]} && \text{(definition)} \\
&= \frac{P[(A \cap B) \cup (C \cap B)]}{P[B]} && \text{(distributive property)} \\
&= \frac{P[A \cap B] + P[C \cap B]}{P[B]} && \begin{array}{l}\text{(Axiom 3 for ordinary probability,} \\ A \cap C = \emptyset \Rightarrow (A \cap B) \cap (C \cap B) = \emptyset)\end{array} \\
&= P[A|B] + P[C|B] && \text{(definition of conditional probability).}
\end{aligned}$$

Conditional probabilities are useful in that they allow us to simplify probability calculations. One particularly important relationship based on conditional probability is described next. Consider a partitioning of the sample space $\mathcal{S}$. Recall that a partition is defined as a group of sets $B_1, B_2, \ldots, B_N$ such that $\mathcal{S} = \cup_{i=1}^N B_i$ and $B_i \cap B_j = \emptyset$ for $i \neq j$. Then we can rewrite the probability $P[A]$ as

$$P[A] = P[A \cap \mathcal{S}] = P\left[A \cap \left(\cup_{i=1}^N B_i\right)\right].$$

But by a slight extension of the distributive property of sets, we can express this as

$$P[A] = P[(A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_N)].$$

Since the $B_i$'s are mutually exclusive, then so are the $A \cap B_i$'s, and therefore

$$P[A] = \sum_{i=1}^N P[A \cap B_i]$$

or finally

$$P[A] = \sum_{i=1}^N P[A|B_i]P[B_i]. \tag{4.4}$$

This relationship is called the *law of total probability*. Its utility is illustrated next.

**Example 4.2 – A compound experiment**

Two urns contain different proportions of red and black balls. Urn 1 has a proportion $p_1$ of red balls and a proportion $1 - p_1$ of black balls whereas urn 2 has proportions of $p_2$ and $1 - p_2$ of red balls and black balls, respectively. A *compound* experiment is performed in which an urn is chosen at random, followed by the selection of a ball. We would like to find the probability that a red ball is selected. To do so we use (4.4) with $A = \{\text{red ball selected}\}$, $B_1 = \{\text{urn 1 chosen}\}$, and $B_2 = \{\text{urn 2 chosen}\}$. Then

$$\begin{aligned} P[\text{red ball selected}] &= P[\text{red ball selected}|\text{urn 1 chosen}]P[\text{urn 1 chosen}] \\ &\quad + P[\text{red ball selected}|\text{urn 2 chosen}]P[\text{urn 2 chosen}] \\ &= p_1\tfrac{1}{2} + p_2\tfrac{1}{2} = \tfrac{1}{2}(p_1 + p_2). \end{aligned}$$

◇

⚠ **Do $B_1$ and $B_2$ really partition the sample space?**

To verify that the application of the law of total probability is indeed valid for this problem, we need to show that $B_1 \cup B_2 = \mathcal{S}$ and $B_1 \cap B_2 = \emptyset$. In our description of $B_1$ and $B_2$ we refer to the choice of an urn. In actuality, this is shorthand for all the balls in the urn. If urn 1 contains balls numbered 1 to $N_1$, then by choosing urn 1 we are really saying that the event is that one of the balls numbered 1 to $N_1$ is chosen and similarly for urn 2 being chosen. Hence, since the sample space consists of all the numbered balls in urns 1 and 2, it is observed that the union of $B_1$ and $B_2$ is the set of all possible outcomes or the sample space. Also, $B_1$ and $B_2$ are mutually exclusive since we choose urn 1 *or* urn 2 but not both.

⚠

Some more examples follow.

**Example 4.3 – Probability of error in a digital communication system**

In a digital communication system a "0" or "1" is transmitted to a receiver. Typically, either *bit* is equally likely to occur so that a *prior probability* of 1/2 is assumed. At the receiver a decoding error can be made due to channel noise, so that a 0 may be mistaken for a 1 and vice versa. Defining the probability of decoding a 1 when a 0 is transmitted as $\epsilon$ and a 0 when a 1 is transmitted also as $\epsilon$, we are interested in the overall probability of an error. A probabilistic model summarizing the relevant features is shown in Figure 4.3. Note that the problem at hand is essentially the same as the previous one. If urn 1 is chosen, then we transmit a 0 and if urn 2 is chosen, we transmit a 1. The effect of the channel is to introduce an error so that even if we know which bit was transmitted, we do not know the received bit. This is analogous to not knowing which ball was chosen from the given urn. The

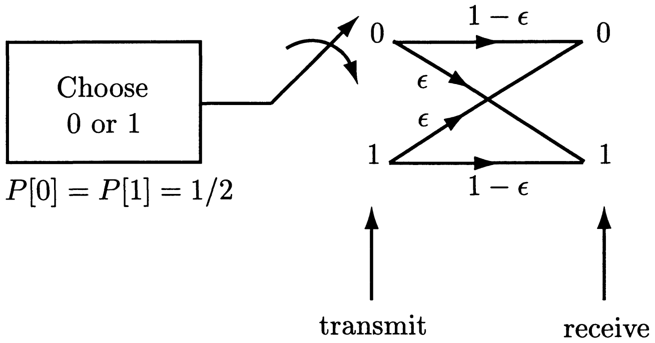


Figure 4.3: Probabilistic model of a digital communication system.

probability of error is from (4.4)

$$\begin{aligned} P[\text{error}] &= P[\text{error}|0 \text{ transmitted}]P[0 \text{ transmitted}] \\ &\quad + P[\text{error}|1 \text{ transmitted}]P[1 \text{ transmitted}] \\ &= \epsilon\tfrac{1}{2} + \epsilon\tfrac{1}{2} = \epsilon. \end{aligned}$$

◇

Conditional probabilities can be quite tricky, in that they sometimes produce counterintuitive results. A famous instance of this is the Monty Hall or Let's Make a Deal problem.

**Example 4.4 – Monty Hall problem**

About 40 years ago there was a television game show called "Let's Make a Deal". The game show host, Monty Hall, would present the contestant with three closed doors. Behind one door was a new car, while the others concealed less desireable prizes, for instance, farm animals. The contestant would first have the opportunity to choose a door, but it would not be opened. Monty would then choose one of the remaining doors and open it. Since he would have knowledge of which door led to the car, he would always choose a door to reveal one of the farm animals. Hence, if the contestant had chosen one of the farm animals, Monty would then choose the door that concealed the other farm animal. If the contestant had chosen the door behind which was the car, then Monty would choose one of the other doors, both concealing farm animals, at random. At this point in the game, the contestant was faced with two closed doors, one of which led to the car and the other to a farm animal. The contestant was given the option of either opening the door she had originally chosen or deciding to open the other door. What should she do? The answer, surprisingly, is that by choosing to switch doors she has a probability of 2/3 of winning the car! If she stays with her original choice, then the probability is only 1/3. Most people would say that irregardless of which strategy she decided upon, her probability of winning the car is 1/2.

|  |  |  | $M_j$ |  |
|---|---|---|---|---|
|  |  | 1 | 2 | 3 |
|  | 1 | 0 | $\frac{1}{6}$ | $\frac{1}{6}$ |
| $C_i$ | 2 | 0 | 0 | $\frac{1}{3}^*$ |
|  | 3 | 0 | $\frac{1}{3}^*$ | 0 |

Table 4.2: Joint probabilities ($P[C_i, M_j] = P[M_j|C_i]P[C_i]$) for contestant's initial and Monty's choice of doors. Winning door is 1.

To see how these probabilities are determined first assume she stays with her original choice. Then, since the car is equally likely to be placed behind any of the three doors, the probability of the contestant's winning the car is 1/3. Monty's choice of a door is irrelevant since her final choice is always the same as her initial choice. However, if as a result of Monty's action a different door is selected by the contestant, then the probability of winning becomes a *conditional probability.* We now compute this by assuming that the car is behind door one. Define the events $C_i = \{\text{contestant initially chooses door } i\}$ for $i = 1, 2, 3$ and $M_j = \{\text{Monty opens door } j\}$ for $j = 1, 2, 3$. Next we determine the joint probabilities $P[C_i, M_j]$ by using

$$P[C_i, M_j] = P[M_j|C_i]P[C_i].$$

Since the winning door is never chosen by Monty, we have $P[M_1|C_i] = 0$. Also, Monty never opens the door initially chosen by the contestant so that $P[M_i|C_i] = 0$. Then, it is easily verified that

$$\begin{aligned} P[M_2|C_3] &= P[M_3|C_2] = 1 && \text{(contestant chooses losing door)} \\ P[M_3|C_1] &= P[M_2|C_1] = \frac{1}{2} && \text{(contestant chooses winning door)} \end{aligned}$$

and $P[C_i] = 1/3$. The joint probabilities are summarized in Table 4.2. Since the contestant always switches doors, the winning events are $(2, 3)$ (the contestant initially chooses door 2 and Monty chooses door 3) and $(3, 2)$ (the contestant initially chooses door 3 and Monty chooses door 2). As shown in Table 4.2 (the entries with asterisks), the total probability is 2/3. This may be verified directly using

$$\begin{aligned} P[\text{final choice is door 1}] &= P[M_3|C_2]P[C_2] + P[M_2|C_3]P[C_3] \\ &= P[C_2, M_3] + P[C_3, M_2]. \end{aligned}$$

Alternatively, the only way she can *lose* is if she initially chooses door one since she always switches doors. This has a probability of 1/3 and hence her probability of winning is 2/3. In effect, Monty, by eliminating a door, has improved her odds!

◇

## 4.4 Statistically Independent Events

Two events $A$ and $B$ are said to be *statistically independent* (or sometimes just *independent*) if $P[A|B] = P[A]$. If this is true, then

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = P[A]$$

which results in the condition for statistical independence of

$$P[A \cap B] = P[A]P[B]. \tag{4.5}$$

An example is shown in Figure 4.2c. There, the probability of $A$ is unchanged if we know that the outcome is contained in the event $B$. Note, however, that once we know that $B$ has occurred, the outcome could not have been in the uncross-hatched region of $A$ but must be in the cross-hatched region. Knowing that $B$ has occurred does in fact affect the possible outcomes. However, it is the *ratio* of $P[A \cap B]$ to $P[B]$ that remains the same.

**Example 4.5 – Statistical independence does not mean one event does not affect another event.**

If a fair die is tossed, the probability of a 2 or a 3 is $P[A = \{2, 3\}] = 1/3$. Now assume we know that the outcome is an even number or $B = \{2, 4, 6\}$. Recomputing the probability

$$\begin{aligned} P[A|B] &= \frac{P[A \cap B]}{P[B]} = \frac{P[\{2\}]}{P[\{2,4,6\}]} \\ &= \frac{1}{3} = P[A]. \end{aligned}$$

Hence, $A$ and $B$ are independent. Yet, knowledge of $B$ occurring has affected the possible outcomes. In particular, the event $A \cap B = \{2\}$ has half as many elements as $A$, but the reduced sample space $S' = B$ also has half as many elements.

◇

The condition for the event $A$ to be independent of the event $B$ is $P[A \cap B] = P[A]P[B]$. Hence, we need only know the *marginal probabilities* or $P[A], P[B]$ to determine the *joint probability* $P[A \cap B]$. In practice, this property turns out to be very useful. Finally, it is important to observe that statistical independence has a symmetry property, as we might expect. If $A$ is independent of $B$, then $B$ must be independent of $A$ since

$$\begin{aligned} P[B|A] &= \frac{P[B \cap A]}{P[A]} && \text{(definition)} \\ &= \frac{P[A \cap B]}{P[A]} && \text{(commutative property)} \\ &= \frac{P[A]P[B]}{P[A]} && \text{($A$ is independent of $B$)} \\ &= P[B] \end{aligned}$$

and therefore $B$ is independent of $A$. Henceforth, we can say that the events $A$ and $B$ are statistically independent of each other, without further elaboration.

⚠ **Statistically independent events are different than mutually exclusive events.**

If $A$ and $B$ are mutually exclusive and $B$ occurs, then $A$ cannot occur. Thus, $P[A|B] = 0$. If $A$ and $B$ are statistically independent and $B$ occurs, then $P[A|B] = P[A]$. Clearly, the probabilities $P[A|B]$ are only the same if $P[A] = 0$. In general then, the conditions of mutually exclusivity and independence must be different since they lead to different values of $P[A|B]$. A specific example of events that *are* both mutually exclusive and statistically independent is shown in Figure 4.4. Finally, the two conditions produce different relationships, namely

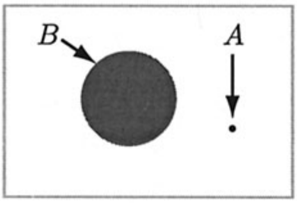


Figure 4.4: Events that are mutually exclusive (since $A \cap B = \emptyset$) and independent (since $P[A \cap B] = P[\emptyset] = 0$ and $P[A]P[B] = 0 \cdot P[B] = 0$).

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] && \text{mutually exclusive events} \\ P[A \cap B] &= P[A]P[B] && \text{statistically independent events.} \end{aligned}$$

See also Figure 4.2c for statistically independent but not mutually exclusive events. Can you think of a case of mutually exclusive but not independent events?

⚠

Consider now the extension of the idea of statistical independence to three events. Three events are defined to be independent if the knowledge that any one or two of the events has occurred does not affect the probability of the third event. For example, one condition is that $P[A|B \cap C] = P[A]$. We will use the shorthand notation $P[A|B, C]$ to indicate that this is the probability of $A$ given that $B$ *and* $C$ has occurred. Note that if $B$ and $C$ has occurred, then by definition $B \cap C$ has occurred. The full set of conditions is

$$\begin{aligned} P[A|B] &= P[A|C] = P[A|B,C] = P[A] \\ P[B|A] &= P[B|C] = P[B|A,C] = P[B] \\ P[C|A] &= P[C|B] = P[C|A,B] = P[C]. \end{aligned}$$

These conditions are satisfied if and only if

$$\begin{aligned} P[AB] &= P[A]P[B] & (4.6)\\ P[AC] &= P[A]P[C] & (4.7)\\ P[BC] &= P[B]P[C] & (4.8)\\ P[ABC] &= P[A]P[B]P[C]. & (4.9) \end{aligned}$$

If the first three conditions (4.6)–(4.8) are satisfied, then the events are said to be *pairwise* independent. They are not enough, however, to ensure independence. The last condition (4.9) is also required since without it we could not assert that

$$\begin{aligned} P[A|B,C] &= P[A|BC] && \text{(definition of } B \text{ and } C \text{ occurring)}\\ &= \frac{P[ABC]}{P[BC]} && \text{(definition of conditional probability)}\\ &= \frac{P[ABC]}{P[B]P[C]} && \text{(from (4.8))}\\ &= \frac{P[A]P[B]P[C]}{P[B]P[C]} && \text{(from (4.9))}\\ &= P[A] \end{aligned}$$

and similarly for the other conditions (see also Problem 4.20 for an example). In general, events $E_1, E_2, \ldots, E_N$ are defined to be statistically independent if

$$\begin{aligned} P[E_iE_j] &= P[E_i]P[E_j] & i \neq j\\ P[E_iE_jE_k] &= P[E_i]P[E_j]P[E_k] & i \neq j \neq k\\ &\vdots \\ P[E_1E_2\cdots E_N] &= P[E_1]P[E_2]\cdots P[E_N]. \end{aligned}$$

Although statistically independent events allow us to compute joint probabilities based on only the marginal probabilities, we can still determine joint probabilities without this property. Of course, it becomes much more difficult. Consider three events as an example. Then, the joint probability is

$$\begin{aligned} P[ABC] &= P[A|B,C]P[BC]\\ &= P[A|B,C]P[B|C]P[C]. & (4.10) \end{aligned}$$

This relationship is called the *probability chain rule.* One is required to determine conditional probabilities, not always an easy matter. A simple example follows.

**Example 4.6 – Tossing a fair die - once again**

If we toss a fair die, then it is clear that the probability of the outcome being 4 is 1/6. We can, however, rederive this result by using (4.10). Letting

$$\begin{aligned} A &= \{\text{even number}\} = \{2,4,6\} \\ B &= \{\text{numbers} > 2\} = \{3,4,5,6\} \\ C &= \{\text{numbers} < 5\} = \{1,2,3,4\} \end{aligned}$$

we have that $ABC = \{4\}$. These events can be shown to be dependent (see Problem 4.21). Now making use of (4.10) and noting that $BC = \{3,4\}$ it follows that

$$\begin{aligned} P[ABC] &= P[A|B,C]P[B|C]P[C] \\ &= \left(\frac{1/6}{2/6}\right)\left(\frac{2/6}{4/6}\right)\left(\frac{4}{6}\right) = \frac{1}{6}. \end{aligned}$$

◇

## 4.5 Bayes' Theorem

The definition of conditional probability leads to a famous and sometimes controversial formula for computing conditional probabilities. Recalling the definition, we have that

$$P[A|B] = \frac{P[AB]}{P[B]} \tag{4.11}$$

and

$$P[B|A] = \frac{P[AB]}{P[A]}. \tag{4.12}$$

Upon substitution of $P[AB]$ from (4.11) into (4.12)

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]}. \tag{4.13}$$

This is called *Bayes' theorem.* By knowing the marginal probabilities $P[A], P[B]$ and the conditional probability $P[A|B]$, we can determine the other conditional probability $P[B|A]$. The theorem allows us to perform "inference" or to assess (with some probability) the validity of an event when some other event has been observed. For example, if an urn containing an unknown composition of balls is sampled with replacement and produces an outcome of 10 red balls, what are we to make of this? One might conclude that the urn contains only red balls. Yet, another individual might claim that the urn is a "fair" one, containing half red balls and half black balls, and attribute the outcome to luck. To test the latter conjecture we now determine the probability of a fair urn given that 10 red balls have just been drawn. The reader should note that we are essentially going "backwards" – usually we compute the probability of choosing 10 red balls *given* a fair urn. Now we are *given* the outcomes and wish to determine the probability of a fair urn. In doing so we believe that the urn is fair with probability 0.9. This is due to our past experience with our purchases from urn.com. In effect, we assume that the prior probability of $B = \{\text{fair urn}\}$ is $P[B] = 0.9$. If $A = \{10 \text{ red balls drawn}\}$, we wish to determine $P[B|A]$, which is the probability of the urn being fair *after* the experiment has been performed or the *posterior* probability. This probability is *our reassessment of the fair urn in light of the new evidence (10 red balls drawn).* Let's compute $P[B|A]$ which according to (4.13) requires knowledge of the *prior probability* $P[B]$ and the conditional probability $P[A|B]$. The former was assumed to be 0.9 and the latter is the probability of drawing 10 successive red balls from an urn with $p = 1/2$. From our previous work this is given by the binomial law as

$$\begin{aligned} P[A|B] &= P[k=10] = \binom{M}{k} p^k (1-p)^{M-k} \\ &= \binom{10}{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^{0} = \left(\frac{1}{2}\right)^{10}. \end{aligned}$$

We still need to find $P[A]$. But this is easily found using the law of total probability as

$$\begin{aligned} P[A] &= P[A|B]P[B] + P[A|B^c]P[B^c] \\ &= P[A|B]P[B] + P[A|B^c](1 - P[B]) \end{aligned}$$

and thus only $P[A|B^c]$ needs to be determined (and which is *not equal to* $1 - P[A|B]$ as is shown in Problem 4.9). This is the conditional probability of drawing 10 red balls from a *unfair urn.* For simplicity we will assume that an unfair urn has all red balls and thus $P[A|B^c] = 1$. Now we have that

$$P[A] = \left(\frac{1}{2}\right)^{10} (0.9) + (1)(0.1)$$

and using this in (4.13) yields

$$P[B|A] = \frac{\left(\frac{1}{2}\right)^{10} (0.9)}{\left(\frac{1}{2}\right)^{10} (0.9) + (1)(0.1)} = 0.0087.$$

The posterior probability (after 10 red balls have been drawn) that the urn is fair is only 0.0087. Our conclusion would be to reject the assumption of a fair urn.

Another way to quantify the result is to compare the posterior probability of the unfair urn to the probability of the fair urn by the ratio of the former to the latter. This is called the *odds ratio* and it is interpreted as the odds *against* the hypothesis of a fair urn. In this case it is

$$\text{odds} = \frac{P[B^c|A]}{P[B|A]} = \frac{1 - 0.0087}{0.0087} = 113.$$

It is seen from this example that based on observed "data", prior beliefs embodied in $P[B] = 0.9$ can be modified to yield posterior beliefs or $P[B|A] = 0.0087$. This is an important concept in statistical inference [Press 2003].

In the previous example, we used the law of total probability to determine the posterior probability. More generally, if a set of $B_i$'s partition the sample space, then Bayes' theorem can be expressed as

$$P[B_k|A] = \frac{P[A|B_k]P[B_k]}{\sum_{i=1}^{N} P[A|B_i]P[B_i]} \qquad k = 1, 2, \ldots, N. \tag{4.14}$$

The denominator in (4.14) serves to normalize the posterior probability so that the conditional probabilities sum to one or

$$\sum_{k=1}^{N} P[B_k|A] = 1.$$

In many problems one is interested in determining whether an observed event or *effect* is the result of some *cause*. Again the backwards or *inferential* reasoning is implicit. Bayes' theorem can be used to quantify this connection as illustrated next.

**Example 4.7 – Medical diagnosis**

Suppose it is known that 0.001% of the general population has a certain type of cancer. A patient visits a doctor complaining of symptoms that might indicate the presence of this cancer. The doctor performs a blood test that will confirm the cancer with a probability of 0.99 if the patient does indeed have cancer. However, the test also produces *false positives* or says a person has cancer when he does not. This occurs with a probability of 0.2. If the test comes back positive, what is the probability that the person has cancer?

To solve this problem we let $B = \{\text{person has cancer}\}$, the causitive event, and $A = \{\text{test is positive}\}$, the effect of that event. Then, the desired probability is

$$\begin{aligned} P[B|A] &= \frac{P[A|B]P[B]}{P[A|B]P[B] + P[A|B^c]P[B^c]} \\ &= \frac{(0.99)(0.00001)}{(0.99)(0.00001) + (0.2)(0.99999)} = 4.95 \times 10^{-5}. \end{aligned}$$

The prior probability of the person having cancer is $P[B] = 10^{-5}$ while the posterior probability of the person having cancer (after the test is performed and found to be positive) is $P[B|A] = 4.95 \times 10^{-5}$. With these results the doctor might be hard pressed to order additional tests. This is quite surprising, and is due to the prior probability assumed, which is quite small and therefore tends to nullify the test results. If we had assumed that $P[B] = 0.5$, for indeed the doctor is seeing a patient who is complaining of symptoms consistent with cancer and not some person chosen at random from the general population, then

$$P[B|A] = \frac{(0.99)(0.5)}{(0.99)(0.5) + (0.2)(0.5)} = 0.83$$

which seems more reasonable (see also Problem 4.23). The controversy surrounding the use of Bayes' theorem in probability calculations can almost always be traced back to the prior probability assumption. Bayes' theorem is mathematically correct – only its application is sometimes in doubt!

◇

## 4.6 Multiple Experiments

### 4.6.1 Independent Subexperiments

An experiment that was discussed in Chapter 1 was the repeated tossing of a coin. We can alternatively view this experiment as a succession of *subexperiments*, with each subexperiment being a single toss of the coin. It is of interest to investigate the relationship between the probabilities defined on the experiment and those defined on the subexperiments. To be more concrete, assume a coin is tossed twice in succession and we wish to determine the probability of the event $A = \{(H,T)\}$. Recall that the notation $(H,T)$ denotes an *ordered* 2-tuple and represents a head on toss 1 and a tail on toss 2. For a fair coin it was determined to be 1/4 since we assumed that all 4 possible outcomes were equally likely. This seemed like a reasonable assumption. However, if the coin had a probability of heads of 0.99, we might not have been so quick to agree with the equally likely assumption. How then are we to determine the probabilities? Let's first consider the experiment to be composed of two separate subexperiments with each subexperiment having a sample space $\mathcal{S}^1 = \{H,T\}$. The sample space of the overall experiment is obtained by forming the *cartesian product*, which for this example is defined as

$$\begin{aligned}\mathcal{S} &= \mathcal{S}^1 \times \mathcal{S}^1 \\ &= \{(i,j) : i \in \mathcal{S}^1; j \in \mathcal{S}^1\} \\ &= \{(H,H),(H,T),(T,H),(T,T)\}.\end{aligned}$$

It is formed by taking an outcome from $\mathcal{S}^1$ for the first element of the 2-tuple and an outcome from $\mathcal{S}^1$ for the second element of the 2-tuple and doing this for all possible outcomes. It would be exceedingly useful if we could determine probabilities for events defined on $\mathcal{S}$ from those probabilities for events defined on $\mathcal{S}^1$. In this way the determination of probabilities of very complicated events could be simplified. Such is the case if we assume that the *subexperiments are independent.* Continuing on, we next calculate $P[A] = P[(H,T)]$ for a coin with an arbitrary probability of heads $p$. This event is defined on the sample space of 2-tuples, which is $\mathcal{S}$. We can, however, express it as an intersection

$$\begin{aligned}\{(H,T)\} &= \{(H,H),(H,T)\} \cap \{(H,T),(T,T)\} \\ &= \{\text{heads on toss } 1\} \cap \{\text{tails on toss } 2\} \\ &= H_1 \cap T_2.\end{aligned}$$

We would expect the events $H_1$ and $T_2$ to be independent of each other. Whether a head or tail appears on the first toss should not affect the probability of the outcome of the second toss and vice versa. Hence, we will let $P[(H,T)] = P[H_1]P[T_2]$ in accordance with the definition of statistically independent events. We can determine $P[H_1]$ either as $P[(H,H),(H,T)]$, which is defined on $\mathcal{S}$ or *equivalently due to the independence assumption* as $P[H]$, which is defined on $\mathcal{S}^1$. Note that $P[H]$ is the marginal probability and is equal to $P[(H,H)] + P[(H,T)]$. But the latter was specified to be $p$ and therefore we have that

$$\begin{aligned}P[H_1] &= p \\ P[T_2] &= 1-p\end{aligned}$$

and finally,

$$P[(H,T)] = p(1-p).$$

For a fair coin we recover the previous value of 1/4, but not otherwise.

Experiments that are composed of subexperiments whose probabilities of the outcomes do not depend on the outcomes of any of the other subexperiments are defined to be *independent subexperiments.* Their utility is to allow calculation of joint probabilities from marginal probabilities. More generally, if we have $M$ independent subexperiments, with $A_i$ an event described for experiment $i$, then the joint event $A = A_1 \cap A_2 \cap \cdots \cap A_M$ has probability

$$P[A] = P[A_1]P[A_2]\cdots P[A_M]. \tag{4.15}$$

Apart from the differences in sample spaces upon which the probabilities are defined, independence of subexperiments is equivalent to statistical independence of events defined on the *same sample space.*

### 4.6.2 Bernoulli Sequence

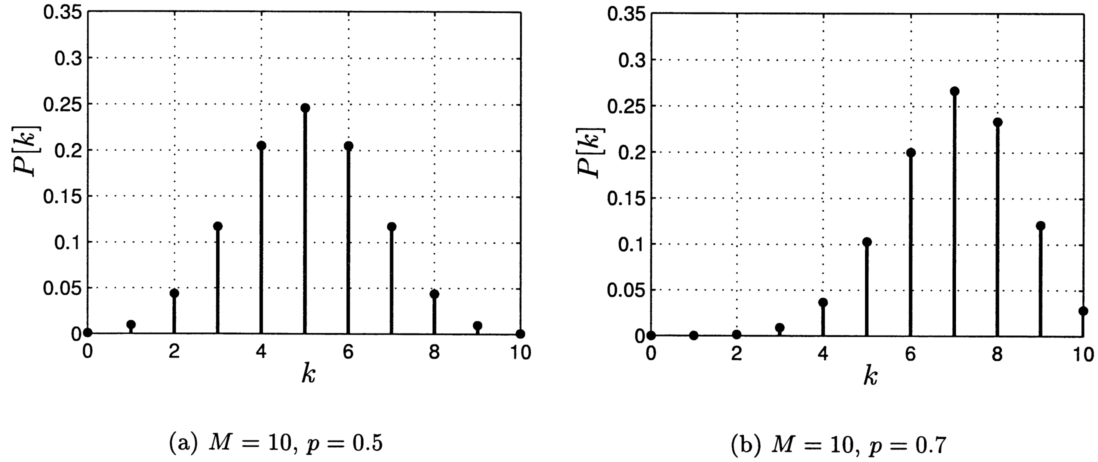
The single tossing of a coin with probability $p$ of heads is an example of a *Bernoulli trial.* Consecutive *independent* Bernoulli trials comprise a *Bernoulli sequence.* More generally, any sequence of $M$ independent subexperiments with each subexperiment producing two possible outcomes is called a Bernoulli sequence. Typically, the subexperiment outcomes are labeled as 0 and 1 with the probability of a 1 being $p$. Hence, for a Bernoulli trial $P[0] = 1-p$ and $P[1] = p$. Several important probability laws are based on this model.

### Binomial Probability Law

Assume that $M$ independent Bernoulli trials are carried out. We wish to determine the probability of $k$ 1's (or successes). Each outcome is an $M$-tuple and a successful outcome would consist of $k$ 1's and $M-k$ 0's in any order. Thus, each successful outcome has a probability of $p^k(1-p)^{M-k}$ due to independence. The total number of successful outcomes is the number of ways $k$ 1's may be placed in the $M$-tuple. This is known from combinatorics to be $\binom{M}{k}$ (see Section 3.8). Hence, by summing up the probabilities of the successful simple events, which are mutually exclusive, we have

$$P[k] = \binom{M}{k} p^k(1-p)^{M-k} \qquad k = 0, 1, \ldots, M \tag{4.16}$$

which we immediately recognize as the binomial probability law. We have previously encountered the same law when we chose $M$ balls at random from an urn with replacement and desired the probability of obtaining $k$ red balls. The proportion of red balls was $p$. In that case, each subexperiment was the choosing of a ball and all the subexperiments were *independent* of each other. The binomial probabilities are shown in Figure 4.5 for various values of $p$.



(a) $M = 10, p = 0.5$

(b) $M = 10, p = 0.7$

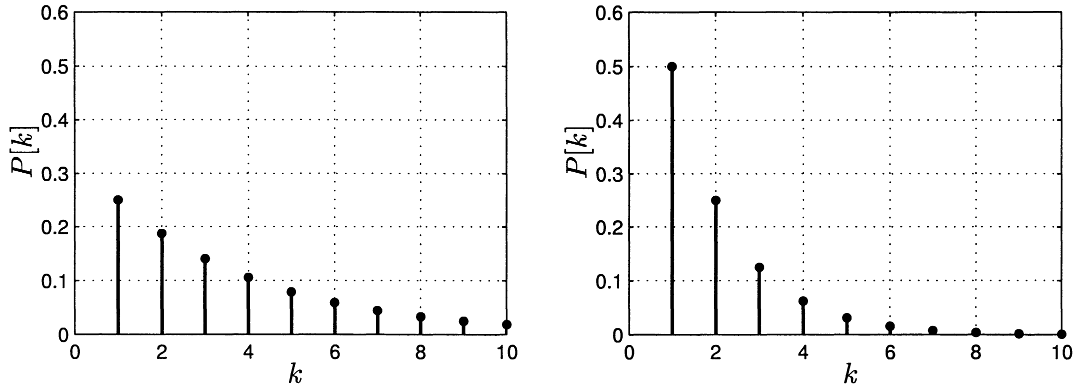
Figure 4.5: The binomial probability law for different values of $p$.

### Geometric Probability Law

Another important aspect of a Bernoulli sequence is the appearance of the first success. If we let $k$ be the Bernoulli trial for which the first success is observed, then the event of interest is the simple event $(\mathrm{f}, \mathrm{f}, \ldots, \mathrm{f}, \mathrm{s})$, where s, f denote success and failure, respectively. This is a $k$-tuple with the first $k-1$ elements all f's. The probability of the first success at trial $k$ is therefore

$$P[k] = (1-p)^{k-1}p \qquad k = 1,2,\ldots \tag{4.17}$$

where $0 < p < 1$. This is called the *geometric probability law*. The geometric probabilities are shown in Figure 4.6 for various values of $p$. It is interesting to note that the first success is always most likely to occur on the first trial or for $k = 1$. This is true even for small values of $p$, which is somewhat counterintuitive. However, upon further reflection, for the first success to occur on trial $k = 1$ we must have a success on trial 1 and the outcomes of the remaining trials are arbitrary. For a success on trial $k = 2$, for example, we must have a failure on trial 1 followed by a success on trial 2, with the remaining outcomes arbitrary. This additional constraint reduces the probability. It will be seen later, though, that the average number of trials required for a success is $1/p$, which is more in line with our intuition. An example of its use follows.



(a) $p = 0.25$

(b) $p = 0.5$

Figure 4.6: The geometric probability law for different values of $p$.

**Example 4.8 – Telephone calling**

A fax machine dials a phone number that is typically busy 80% of the time. The machine dials it every 5 minutes until the line is clear and the fax is able to be transmitted. What is the probability that the fax machine will have to dial the number 9 times? The number of times the line is busy can be considered the number of failures with each failure having a probability of $1 - p = 0.8$. If the number is dialed 9 times, then the first success occurs for $k = 9$ and

$$P[9] = (0.8)^8(0.2) = 0.0336.$$

◊

A useful property of the geometric probability law is that it is memoryless. Assume it is known that no successes occurred in the first $m$ trials. Then, the probability of the first success at trial $m+l$ is the same as if we had started the Bernoulli sequence experiment over again and determined the probability of the first success at trial $l$ (see Problem 4.34).

### 4.6.3 Multinomial Probability Law

Consider an extension to the Bernoulli sequence in which the trials are still independent but the outcomes for each trial may take on more than two values. For example, let $\mathcal{S}^1 = \{1, 2, 3\}$ and denote the probabilities of the outcomes 1, 2, and 3 by $p_1$, $p_2$, and $p_3$, respectively. As usual, the assignment of these probabilities must satisfy $\sum_{i=1}^{3} p_i = 1$. Also, let the number of trials be $M = 6$ so that a possible outcome might be $(2, 1, 3, 1, 2, 2)$, whose probability is $p_2p_1p_3p_1p_2p_2 = p_1^2p_2^3p_3^1$. The multinomial probability law specifies the probability of obtaining $k_1$ 1's, $k_2$ 2's, and $k_3$ 3's, where $k_1 + k_2 + k_3 = M = 6$. In the current example, $k_1 = 2$, $k_2 = 3$, and $k_3 = 1$. Some outcomes with the same number of 1's, 2's', and 3's are $(2, 1, 3, 1, 2, 2)$, $(1, 2, 3, 1, 2, 2)$, $(1, 2, 1, 2, 2, 3)$, etc., with each outcome having a probability of $p_1^2p_2^3p_3^1$. The total number of these outcomes will be the total number of distinct 6-tuples that can be made with the numbers $1, 1, 2, 2, 2, 3$. If the numbers to be used were all different, then the total number of 6-tuples would be 6! , or all permutations. However, since they are not, some of these permutations will be the same. For example, we can arrange the 2's 3! ways and still have the same 6-tuple. Likewise, the 1's can be arranged 2! ways without changing the 6-tuple. As a result, the total number of *distinct* 6-tuples is

$$\frac{6!}{2!3!1!} \tag{4.18}$$

which is called the *multinomial coefficient.* (See also Problem 4.36 for another way to derive this.) It is sometimes denoted by

$$\binom{6}{2,3,1}.$$

Finally, for our example the probability of the sequence exhibiting two 1's, three 2's, and one 3 is

$$\frac{6!}{2!3!1!}p_1^2p_2^3p_3^1.$$

This can be generalized to the case of $M$ trials with $N$ possible outcomes for each trial. The probability of $k_1$ 1's, $k_2$ 2's,..., $k_N$ $N$'s is

$$P[k_1, k_2, \ldots, k_N] = \binom{M}{k_1, k_2, \ldots, k_N} p_1^{k_1}p_2^{k_2}\cdots p_N^{k_N} \qquad k_1 + k_2 + \cdots + k_N = M \tag{4.19}$$

and where $\sum_{i=1}^{N} p_i = 1$. This is termed the *multinomial probability law.* Note that if $N = 2$, then it reduces to the binomial law (see Problem 4.37). An example follows.

**Example 4.9 – A version of scrabble**

A person chooses 9 letters at random from the English alphabet with replacement. What is the probability that she will be able to make the word "committee"? Here we have that the outcome on each trial is one of 26 letters. To be able to make the word she needs $k_c = 1, k_e = 2, k_i = 1, k_m = 2, k_o = 1, k_t = 2$, and $k_{\text{other}} = 0$. We have denoted the outcomes as $c, e, i, m, o, t$, and "other". "Other" represents the remaining 20 letters so that $N = 7$. Thus, the probability is from (4.19)

$$P[k_c = 1, k_e = 2, k_i = 1, k_m = 2, k_o = 1, k_t = 2, k_{\text{other}} = 0] =$$

$$\binom{9}{1,2,1,2,1,2,0} \left(\frac{1}{26}\right)^9 \left(\frac{20}{26}\right)^0$$

since $p_c = p_e = p_i = p_m = p_o = p_t = 1/26$ and $p_{\text{other}} = 20/26$ due to the assumption of "at random" sampling and with replacement. This becomes

$$P[k_c = 1, k_e = 2, k_i = 1, k_m = 2, k_o = 1, k_t = 2, k_{\text{other}} = 0] =$$

$$\frac{9!}{1!2!1!2!1!2!0!} \left(\frac{1}{26}\right)^9 = 8.35 \times 10^{-9}.$$

◇

### 4.6.4 Nonindependent Subexperiments

When the subexperiments are independent, the calculation of probabilities can be greatly simplified. An event that can be written as $A = A_1 \cap A_2 \cap \cdots \cap A_M$ can be found via

$$P[A] = P[A_1]P[A_2] \cdots P[A_M]$$

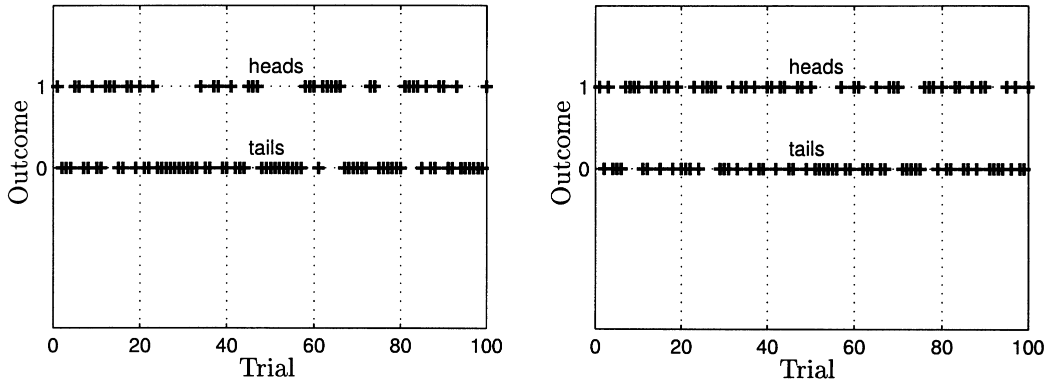
where each $P[A_i]$ can be found by considering only the individual subexperiment. However, the assumption of independence can sometimes be unreasonable. In the absence of independence, the probability would be found by using the chain rule (see (4.10) for $M = 3$)

$$P[A] = P[A_M|A_{M-1}, \ldots, A_1]P[A_{M-1}|A_{M-2}, \ldots, A_1] \cdots P[A_2|A_1]P[A_1]. \quad (4.20)$$

Such would be the case if a Bernoulli sequence were composed of nonindependent trials as illustrated next.

**Example 4.10 – Dependent Bernoulli trials**

Assume that we have two coins. One is fair and the other is weighted to have a probability of heads of $p \neq 1/2$. We begin the experiment by first choosing at random one of the two coins and then tossing it. If it comes up heads, we choose the fair coin to use on the next trial. If it comes up tails, we choose the weighted coin to use on the next trial. We repeat this procedure for all the succeeding trials. One possible sequence of outcomes is shown in Figure 4.7a for the weighted coin having $p = 1/4$. Also shown is the case when $p = 1/2$ or a fair coin is always used, so that we are equally likely to observe a head or a tail on each trial. Note that in the case of $p = 1/4$ (see Figure 4.7a), if the outcome is a tail on any trial, then we use the weighted coin for the next trial. Since the weighted coin is biased towards producing a tail, we would expect to again see a tail, and so on. This accounts for the long run of tails observed. Clearly, the trials are not independent.



(a) $M = 100$, $p = 0.25$

(b) $M = 100$, $p = 0.5$

Figure 4.7: Dependent Bernoulli sequence for different values of $p$.

◇

If we think some more about the previous experiment, we realize that the dependency between trials is due only to the outcome of the $(i-1)^{\text{st}}$ trial affecting the outcome of the $i$th trial. In fact, once the coin has been chosen, the probabilities for the next trial are either $P[0] = P[1] = 1/2$ if a head occurred on the previous trial or $P[0] = 3/4, P[1] = 1/4$ if the previous trial produced a tail. The previous outcome is called the *state* of the sequence. This behavior may be summarized by the *state probability diagram* shown in Figure 4.8. The probabilities shown are actually conditional probabilities. For example, 3/4 is the probability $P[\text{tail on } i\text{th toss}|\text{tail on } i-1^{\text{st}} \text{ toss}] = P[0|0]$, and similarly for the others. This type of Bernoulli sequence, in which the probabilities for trial $i$ depend only on the outcome of the previous trial, is called a *Markov sequence.* Mathematically, the probability of the event $A_i$ on the $i$th trial given all the previous outcomes can be written as

$$P[A_i|A_{i-1}, A_{i-2}, \ldots, A_1] = P[A_i|A_{i-1}].$$

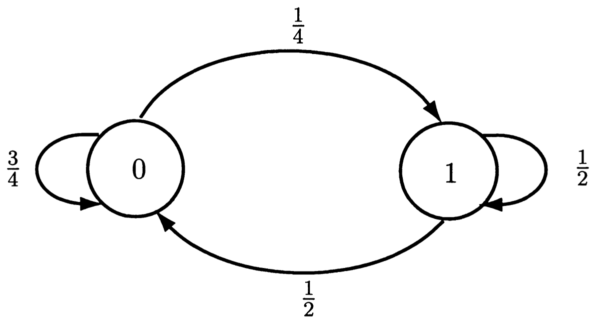


Figure 4.8: Markov state probability diagram.

Using this in (4.20) produces

$$P[A] = P[A_M|A_{M-1}]P[A_{M-1}|A_{M-2}]\cdots P[A_2|A_1]P[A_1]. \tag{4.21}$$

The conditional probabilities $P[A_i|A_{i-1}]$ are called the *state transition probabilities*, and along with the initial probability $P[A_1]$, the probability of any joint event can be determined. For example, we might wish to determine the probability of $N = 10$ tails in succession or of the event $A = \{(0,0,0,0,0,0,0,0,0,0)\}$. If the weighted coin was actually fair, then $P[A] = (1/2)^{10} = 0.000976$, but if $p = 1/4$, we have by letting $A_i = \{0\}$ for $i = 1,2,\ldots,10$ in (4.21)

$$P[A] = \left(\prod_{i=2}^{10} P[A_i|A_{i-1}]\right) P[A_1].$$

But $P[A_i|A_{i-1}] = P[0|0] = P[\text{tails}|\text{weighted coin}] = 3/4$ for $i = 2,3,\ldots,10$. Since we initially choose one of the coins at random, we have

$$\begin{aligned} P[A_1] &= P[0] = P[\text{tail}|\text{weighted coin}]P[\text{weighted coin}] \\ &\quad + P[\text{tail}|\text{fair coin}]P[\text{fair coin}] \\ &= \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{5}{8}. \end{aligned}$$

Thus, we have that

$$P[A] = \left(\prod_{i=2}^{10} \frac{3}{4}\right)\left(\frac{5}{8}\right) = 0.0469$$

or about 48 times more probable than if the weighted coin were actually fair. Note that we could also represent the process by using a *trellis diagram* as shown in Figure 4.9. The probability of any sequence is found by tracing the sequence values through the trellis and multiplying the probabilities for each branch together, along with the initial probability. Referring to Figure 4.9 the sequence $1,0,0$ has a probability of $(3/8)(1/2)(3/4)$. The foregoing example is a simple case of a *Markov chain*. We will study this modeling in much more detail in Chapter 22.

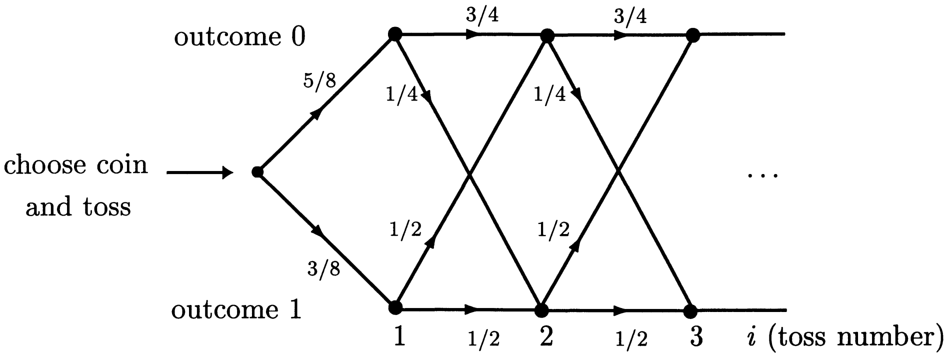


Figure 4.9: Trellis diagram.

## 4.7 Real-World Example – Cluster Recognition

In many areas an important problem is the detection of a "cluster." Epidemiology is concerned with the incidence of a greater than expected number of disease cases in a given geographic area. If such a situation is found to exist, then it may indicate a problem with the local water supply, as an example. Police departments may wish to focus their resources on areas of a city that exhibit an unusually high incidence of crime. Portions of a remotely sensed image may exhibit an increased number of noise bursts. This could be due to a group of sensors that are driven by a faulty power source. In all these examples, we wish to determine if a cluster of events has occurred. By cluster, we mean that more occurrences of an event are observed than would normally be expected. An example could be a geographic area which is divided into a grid of 50 × 50 cells as shown in Figure 4.10. It is seen that an event or "hit", which is denoted by a black square, occurs rather infrequently. In this example, it occurs $29/2500 = 1.16\%$ of the time. Now consider Figure 4.11. We see that the shaded area appears to exhibit more hits than the expected $145 \times 0.0116 = 1.68$ number. One might be inclined to call this shaded area a cluster. But how probable is this cluster? And how can we make a decision to either accept the hypothesis that this area is a cluster or to reject it? To arrive at a decision we use a Bayesian approach. It computes the *odds ratio against* the occurrence of a cluster (or in favor of no cluster), which is defined as

$$\text{odds} = \frac{P[\text{no cluster}|\text{observed data}]}{P[\text{cluster}|\text{observed data}]}.$$

If this number is large, typically much greater than one, we would be inclined to reject the hypothesis of a cluster, and otherwise, to accept it. We can use Bayes' theorem to evaluate the odds ratio by letting $B = \{\text{cluster}\}$ and $A = \{\text{observed data}\}$. Then,

$$\text{odds} = \frac{P[B^c|A]}{P[B|A]} = \frac{P[A|B^c]P[B^c]}{P[A|B]P[B]}.$$

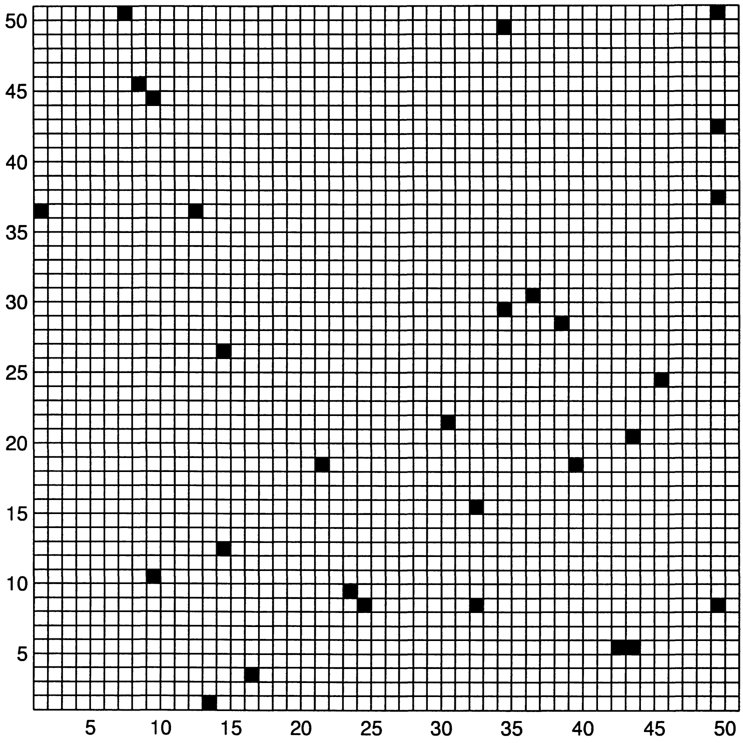


Figure 4.10: Geographic area with incidents shown as black squares – no cluster present.

Note that $P[A]$ is not needed since it cancel outs in the ratio. To evaluate this we need to determine $P[B], P[A|B^c], P[A|B]$. The first probability $P[B]$ is the prior probability of a cluster. Since we believe a cluster is quite unlikely, we assign a probability of $10^{-6}$ to this. Next we need $P[A|B^c]$ or the probability of the observed data if there is no cluster. Since each cell can take on only one of two values, either a hit or no hit, and if we assume that the outcomes of the various cells are independent of each other, we can model the data as a Bernoulli sequence. For this problem, we might be tempted to call it a Bernoulli *array* but the determination of the probabilities will of course proceed as usual. If $M$ cells are contained in the supposed cluster area (shown as shaded in Figure 4.11 with $M = 145$), then the probability of $k$ hits is given by the binomial law

$$P[k] = \binom{M}{k} p^k(1-p)^{M-k}.$$

Next must assign values to $p$ under the hypothesis of a cluster present and no cluster present. From Figure 4.10 in which we did not suspect a cluster, the relative

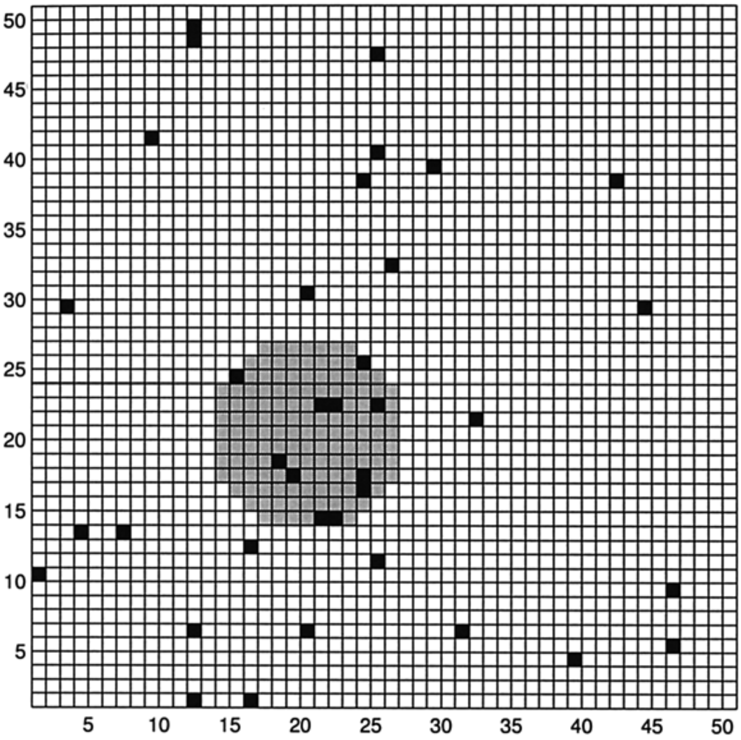


Figure 4.11: Geographic area with incidents shown as black squares – possible cluster present.

frequency of hits was about 0.0116 so that we assume $p_{\mathrm{nc}} = 0.01$ when there is no cluster. When we believe a cluster is present, we assume that $p_{\mathrm{c}} = 0.1$ in accordance with the relative frequency of hits in the shaded area of Figure 4.11, which is 11/145=0.07. Thus,

$$\begin{aligned}
P[A|B^c] &= P[\text{observed data}|\text{no cluster}] = \binom{M}{k} p_{\mathrm{nc}}^k (1-p_{\mathrm{nc}})^{M-k} \\
&= P[k=11|\text{no cluster}] = \binom{145}{11} (0.01)^{11}(0.99)^{134} \\
P[A|B] &= P[\text{observed data}|\text{cluster}] = \binom{M}{k} p_{\mathrm{c}}^k (1-p_{\mathrm{c}})^{M-k} \\
&= P[k=11|\text{cluster}] = \binom{145}{11} (0.1)^{11}(0.9)^{134}
\end{aligned}$$

which results in an odds ratio of

$$\text{odds} = \frac{(0.01)^{11}(0.99)^{134}(1-10^{-6})}{(0.1)^{11}(0.9)^{134}(10^{-6})} = 3.52.$$

Since the posterior probability of no cluster is 3.52 times larger than the posterior probability of a cluster, we would reject the hypothesis of a cluster present. However, the odds against a cluster being present are not overwhelming. In fact, the computer simulation used to generate Figures 4.11 employed $p = 0.01$ for the unshaded region and $p = 0.1$ for the shaded cluster region. The reader should be aware that it is mainly the influence of the small prior probability of a cluster, $P[B] = 10^{-6}$, that has resulted in the greater than unity odds ratio and a decision to reject the cluster present hypothesis.

## References

S. Press, *Subjective and Objective Bayesian Statistics*, John Wiley & Sons, New York, 2003.

D. Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W.H. Freeman, New York, 2001.

## Problems

**4.1 (f)** If $B \subset A$, what is $P[A|B]$? Explain your answer.

**4.2 (☺) (f)** A point $x$ is chosen at random within the interval $(0, 1)$. If it is known that $x \geq 1/2$, what is the probability that $x \geq 7/8$?

**4.3 (w)** A coin is tossed three times with each 3-tuple outcome being equally likely. Find the probability of obtaining $(H, T, H)$ if it is known that the outcome has 2 heads. Do this by 1) using the idea of a reduced sample space and 2) using the definition of conditional probability.

**4.4 (w)** Two dice are tossed. Each 2-tuple outcome is equally likely. Find the probability that the number that comes up on die 1 is the same as the number that comes up on die 2 if it is known that the sum of these numbers is even.

**4.5 (☺) (f)** An urn contains 3 red balls and 2 black balls. If two balls are chosen without replacement, find the probability that the second ball is black if it is known that the first ball chosen is black.

**4.6 (f)** A coin is tossed 11 times in succession. Each 11-tuple outcome is equally likely to occur. If the first 10 tosses produced all heads, what is the probability that the 11th toss will also be a head?

**4.7 (☺) (w)** Using Table 4.1, determine the probability that a college student will have a weight greater than 190 lbs. if he/she has a height exceeding $5'8''$. Next, find the probability that a student's weight will exceed 190 lbs.

**4.8 (w)** Using Table 4.1, find the probability that a student has weight less than 160 lbs. if he/she has height *greater* than $5'4''$. Also, find the probability that a student's weight is less than 160 lbs. if he/she has height *less* than $5'4''$. Are these two results related?

**4.9 (t)** Show that the statement $P[A|B] + P[A|B^c] = 1$ is false. Use Figure 4.2a to provide a counterexample.

**4.10 (t)** Prove that for the events $A, B, C$, which are not necessarily mutually exclusive,

$$P[A \cup B|C] = P[A|C] + P[B|C] - P[AB|C].$$

**4.11 (☺) (w)** A group of 20 patients afflicted with a disease agree to be part of a clinical drug trial. The group is divided up into two groups of 10 subjects each, with one group given the drug and the other group given sugar water, i.e., this is the control group. The drug is 80% effective in curing the disease. If one is not given the drug, there is still a 20% chance of a cure due to remission. What is the probability that a randomly selected subject will be cured?

**4.12 (w)** A new bus runs on Sunday, Tuesday, Thursday, and Saturday while an older bus runs on the other days. The new bus has a probability of being on time of 2/3 while the older bus has a probability of only 1/3. If a passenger chooses an arbitrary day of the week to ride the bus, what is the probability that the bus will be on time?

**4.13 (w)** A digital communication system transmits one of the three values $-1, 0, 1$. A channel adds noise to cause the decoder to sometimes make an error. The error rates are 12.5% if a $-1$ is transmitted, 75% if a 0 is transmitted, and 12.5% if a 1 is transmitted. If the probabilities for the various symbols being transmitted are $P[-1] = P[1] = 1/4$ and $P[0] = 1/2$, find the probability of error. Repeat the problem if $P[-1] = P[0] = P[1]$ and explain your results.

**4.14 (☺) (w)** A sample space is given by $\mathcal{S} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Determine $P[A|B]$ for the events

$$\begin{aligned} A &= \{(x,y) : y \leq 2x, 0 \leq x \leq 1/2 \text{ and } y \leq 2 - 2x, 1/2 \leq x \leq 1\} \\ B &= \{(x,y) : 1/2 \leq x \leq 1, 0 \leq y \leq 1\}. \end{aligned}$$

Are $A$ and $B$ independent?

**4.15 (w)** A sample space is given by $\mathcal{S} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Are the events

$$\begin{aligned} A &= \{(x,y) : y \leq x\} \\ B &= \{(x,y) : y \leq 1 - x\} \end{aligned}$$

independent? Repeat if $B = \{(x, y) : x \leq 1/4\}$.

**4.16 (t)** Give an example of two events that are mutually exclusive but not independent. Hint: See Figure 4.4.

**4.17 (t)** Consider the sample space $\mathcal{S} = \{(x,y,z) : 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\}$, which is the unit cube. Can you find three events that are independent? Hint: See Figure 4.2c.

**4.18 (t)** Show that if (4.9) is satisfied for *all* possible events, then pairwise independence follows. In this case all events are independent.

**4.19 (☺) (f)** It is known that if it rains, there is a 50% chance that a sewer will overflow. Also, if the sewer overflows, then there is a 30% chance that the road will flood. If there is a 20% chance that it will rain, what is the probability that the road will flood?

**4.20 (w)** Consider the sample space $\mathcal{S} = \{1,2,3,4\}$. Each simple event is equally likely. If $A = \{1,2\}, B = \{1,3\}, C = \{1,4\}$, are these events pairwise independent? Are they independent?

**4.21 (☺) (w)** In Example 4.6 determine if the events are pairwise independent. Are they independent?

**4.22 (☺) (w)** An urn contains 4 red balls and 2 black balls. Two balls are chosen in succession without replacement. If it is known that the first ball drawn is black, what are the odds in favor of a red ball being chosen on the second draw?

**4.23 (w)** In Example 4.7 plot the probability that the person has cancer given that the test results are positive, i.e., the posterior probability, as a function of the prior probability $P[B]$. How is the posterior probability that the person has cancer related to the prior probability?

**4.24 (w)** An experiment consists of two subexperiments. First a number is chosen at random from the interval $(0,1)$. Then, a second number is chosen at random from the same interval. Determine the sample space $\mathcal{S}^2$ for the overall experiment. Next consider the event $A = \{(x,y) : 1/4 \leq x \leq 1/2, 1/2 \leq y \leq 3/4\}$ and find $P[A]$. Relate $P[A]$ to the probabilities defined on $S^1 = \{u : 0 < u < 1\}$, where $\mathcal{S}^1$ is the sample space for each subexperiment.

**4.25 (w,c)** A fair coin is tossed 10 times. What is the probability of a run of exactly 5 heads in a row? Do not count runs of 6 or more heads in a row. Now verify your solution using a computer simulation.

**4.26 (☺) (w)** A lady claims that she can tell whether a cup of tea containing milk had the tea poured first or the milk poured first. To test her claim an experiment is set up whereby at random the milk or tea is added first to an empty cup. This experiment is repeated 10 times. If she correctly identifies which liquid was poured first 8 times out of 10, how likely is it that she is guessing? See [Salsburg 2001] for a further discussion of this famous problem.

**4.27 (f)** The probability $P[k]$ is given by the binomial law. If $M = 10$, for what value of $p$ is $P[3]$ maximum? Explain your answer.

**4.28 (⌣̈) (f)** A sequence of independent subexperiments is conducted. Each subexperiment has the outcomes "success", "failure", or "don't know". If $P[\text{success}] = 1/2$ and $P[\text{failure}] = 1/4$, what is the probability of 3 successes in 5 trials?

**4.29 (c)** Verify your results in Problem 4.28 by using a computer simulation.

**4.30 (w)** A drunk person wanders aimlessly along a path by going forward one step with probability 1/2 and going backward one step with probability 1/2. After 10 steps what is the probability that he has moved 2 steps forward?

**4.31 (f)** Prove that the geometric probability law (4.17) is a valid probability assignment.

**4.32 (w)** For a sequence of independent Bernoulli trials find the probability of the first failure at the $k$th trial for $k = 1, 2, \ldots$.

**4.33 (⌣̈) (w)** For a sequence of independent Bernoulli trials find the probability of the second success occurring at the $k$th trial.

**4.34 (t)** Consider a sequence of independent Bernoulli trials. If it is known that the first $m$ trials resulted in failures, prove that the probability of the first success occurring at $m + l$ is given by the geometric law with $k$ replaced by $l$. In other words, the probability is the same as if we had started the process over again after the $m$th failure. There is no memory of the first $m$ failures.

**4.35 (f)** An urn contains red, black, and white balls. The proportion of red is 0.4, the proportion of black is 0.4, and the proportion of white is 0.2. If 5 balls are drawn with replacement, what is the probability of 2 red, 2 black, and 1 white in any order?

**4.36 (t)** We derive the multinomial coefficient for $N = 3$. This will yield the number of ways that an $M$-tuple can be formed using $k_1$ 1's, $k_2$ 2's and $k_3$ 3's. To do so choose $k_1$ places in the $M$-tuple for the 1's. There will be $M - k_1$ positions remaining. Of these positions choose $k_2$ places for the 2's. Fill in the remaining $k_3 = M - k_1 - k_2$ positions using the 3's. Using this result, determine the number of different $M$ digit sequences with $k_1$ 1's, $k_2$ 2's, and $k_3$ 3's.

**4.37 (t)** Show that the multinomial probability law reduces to the binomial law for $N = 2$.

**4.38** (☺) **(w,c)** An urn contains 3 red balls, 3 black balls, and 3 white balls. If 6 balls are chosen with replacement, how many of each color is most likely? Hint: You will need a computer to evaluate the probabilities.

**4.39** **(w,c)** For the problem discussed in Example 4.10 change the probability of heads for the weighted coin from $p = 0.25$ to $p = 0.1$. Redraw the Markov state probability diagram. Next, using a computer simulation generate a sequence of length 100. Explain your results.

**4.40** (☺) **(f)** For the Markov state diagram shown in Figure 4.8 with an initial state probability of $P[0] = 3/4$, find the probability of the sequence $0, 1, 1, 0$.

**4.41** **(f)** A *two-state* Markov chain (see Figure 4.8) has the *state transition probabilities* $P[0|0] = 1/4, P[0|1] = 3/4$ and the initial state probability of $P[0] = 1/2$. What is the probability of the sequence $0, 1, 0, 1, 0$?

**4.42** **(w)** A digital communication system model is shown in Figure 4.12. It consists of two sections with each one modeling a different portion of the communication channel. What is the probability of a bit error? Compare this to the probability of error for the single section model shown in Figure 4.3, assuming that $\epsilon < 1/2$, which is true in practice. Note that Figure 4.12 is a trellis.

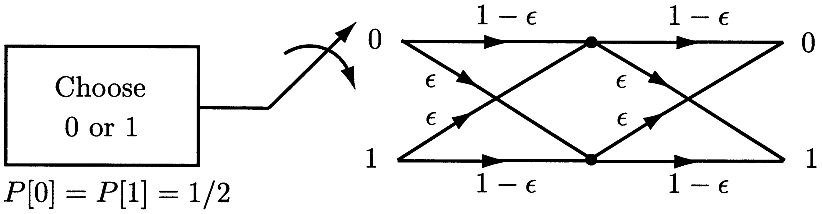


Figure 4.12: Probabilistic model of a digital communication system with two sections.

**4.43** (☺) **(f)** For the trellis shown in Figure 4.9 find the probability of the event $A = \{(0, 1, 0, 0), (0, 0, 0, 0)\}$.