Chapter 1

Introduction

1.1 MOTIVATION

Computer scientists and engineers need powerful techniques to analyze algorithms and computer systems. Similarly, networking engineers need methods to analyze the behavior of protocols, routing algorithms, and congestion in networks. Computer systems and networks are subject to failure, and hence methods for their reliability and availability are needed. Many of the tools necessary for these analyses have their foundations in probability theory. For example, in the analysis of algorithm execution times, it is common to draw a distinction between the *worst-case* and the *average-case* behavior of an algorithm. The distinction is based on the fact that for certain problems, while an algorithm may require an inordinately long time to solve the least favorable instance of the problem, the average solution time is considerably shorter. When many instances of a problem have to be solved, the probabilistic (or average-case) analysis of the algorithm is likely to be more useful. Such an analysis accounts for the fact that the performance of an algorithm is dependent on the distributions of input data items. Of course, we have to specify the relevant probability distributions before the analysis can be carried out. Thus, for instance, while analyzing a sorting algorithm, a common assumption is that every permutation of the input sequence is equally likely to occur.

Similarly, if the storage is dynamically allocated, a probabilistic analysis of the storage requirement is more appropriate than a worst-case analysis. In a like fashion, a worst-case analysis of the accumulation of roundoff errors in a numerical algorithm tends to be rather pessimistic; a probabilistic analysis, although harder, is more useful.

When we consider the analysis of a Web server serving a large number of users, several types of random phenomena need to be accounted for. First, the arrival pattern of requests is subject to randomness due to a large population of diverse users. Second, the resource requirements of requests will likely fluctuate from request to request as well as during the execution of a single request. Finally, the resources of the Web server are subject to random failures due to environmental conditions and aging phenomena. The theory of stochastic (random) processes is very useful in evaluating various measures of system effectiveness such as throughput, response time, reliability, and availability.

Before an algorithm (or protocol) or a system can be analyzed, various probability distributions have to be specified. Where do the distributions come from? We may collect data during the actual operation of the system (or the algorithm). These measurements can be performed by hardware monitors, software monitors, or both. Such data must be analyzed and compressed to obtain the necessary distributions that drive the analytical models discussed above. Mathematical statistics provides us with techniques for this purpose, such as the **design of experiments**, **hypothesis testing**, **estimation**, **analysis of variance**, and **linear and nonlinear regression**.

1.2 PROBABILITY MODELS

Probability theory is concerned with the study of random (or chance) phenomena. Such phenomena are characterized by the fact that their future behavior is not predictable in a deterministic fashion. Nevertheless, such phenomena are usually capable of mathematical descriptions due to certain statistical regularities. This can be accomplished by constructing an idealized probabilistic model of the real-world situation. Such a model consists of a list of all possible outcomes and an assignment of their respective probabilities. The theory of probability then allows us to predict or deduce patterns of future outcomes.

Since a model is an abstraction of the real-world problem, predictions based on the model must be validated against actual measurements collected from the real phenomena. A poor validation may suggest modifications to the original model. The theory of statistics facilitates the process of validation. Statistics is concerned with the inductive process of drawing inferences about the model and its parameters based on the limited information contained in real data.

The role of probability theory is to analyze the behavior of a system or an algorithm assuming the given probability assignments and distributions. The results of this analysis are as good as the underlying assumptions. Statistics helps us in choosing these probability assignments and in the process of validating model assumptions. The behavior of the system (or the algorithm) is observed, and an attempt is made to draw inferences about the underlying unknown distributions of random variables that describe system activity. Methods of statistics, in turn, make heavy use of probability theory.

Consider the problem of predicting the number of request arrivals to a Web server in a fixed time interval $(0,t]$. A common model of this situation is to assume that the number of arrivals in this period has a particular distribution, such as the Poisson distribution (see Chapter 2). Thus we have replaced a complex physical situation by a simple model with a single unknown parameter, namely, the average arrival rate λ. With the help of probability theory we can then deduce the pattern of future arrivals. On the other hand, statistical techniques help us estimate the unknown parameter λ based on actual observations of past arrival patterns. Statistical techniques also allow us to test the validity of the Poisson model.

As another example, consider a fault-tolerant computer system with automatic error recovery capability. Model this situation as follows. The probability of successful recovery is c and probability of an abortive error is $1 - c$. The uncertainty of the physical situation is once again reduced to a simple probability model with a single unknown parameter c. In order to estimate parameter c in this model, we observe N errors out of which n are successfully recovered. A reasonable estimate of c is the relative frequency n/N, since we expect this ratio to converge to c in the limit $N \to \infty$. Note that this limit is a limit in a probabilistic sense:

$$\lim_{N\to\infty} P\left(\left|\frac{n}{N} - c\right| > \epsilon\right) = 0.$$

Axiomatic approaches to probability allow us to define such limits in a mathematically consistent fashion (e.g., see the law of large numbers in Chapter 4) and hence allow us to use relative frequencies as estimates of probabilities.

1.3 SAMPLE SPACE

Probability theory is rooted in the real-life situation where a person performs an experiment the outcome of which may not be certain. Such an experiment is called a **random experiment**. Thus, an experiment may consist of the simple process of noting whether a component is functioning properly or has failed; it may consist of determining the execution time of a program; or it may consist of determining the response time of a server request. The result of any such observations, whether they are simple "yes" or "no" answers, meter readings, or whatever, are called **outcomes** of the experiment.

Definition (Sample Space). The totality of the possible outcomes of a random experiment is called the **sample space** of the experiment and it will be denoted by the letter S.

We point out that the sample space is not determined completely by the experiment. It is partially determined by the purpose for which the experiment is carried out. If the status of two components is observed, for some purposes it is sufficient to consider only three possible outcomes: two functioning, two malfunctioning, one functioning and one malfunctioning. These three outcomes constitute the sample space S. On the other hand, we might be interested in exactly which of the components has failed, if any has failed. In this case the sample space S must be considered as four possible outcomes where the earlier single outcome of one failed, one functioning is split into two outcomes: first failed, second functioning and first functioning, second failed.

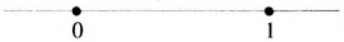


Figure 1.1. A one-dimensional sample space

Many other sample spaces can be defined if we take into account such things as type of failure and so on.

Frequently, we use a larger sample space than is strictly necessary because it is easier to use; specifically, it is always easier to discard excess information than to recover lost information. For instance, in the preceding illustration, the first sample space might be denoted $S_1 = \{0, 1, 2\}$ (where each number indicates how many components are functioning) and the second sample space might be denoted $S_2 = \{(0,0), (0,1), (1,0), (1,1)\}$ (where 0 = failed, 1 = functioning). Given a selection from S_2, we can always add the two components to determine the corresponding choice from S_1; but, given a choice from S_1 (in particular 1), we cannot necessarily recover the corresponding choice from S_2.

It is useful to think of the outcomes of an experiment, the **elements** of the sample space, as points in a space of one or more dimensions. For example, if an experiment consists of examining the state of a single component, it may be functioning properly (denoted by the number 1), or it may have failed (denoted by the number 0). The sample space is one-dimensional, as shown in Figure 1.1. If a system consists of two components there are four possible outcomes, as shown in the two-dimensional sample space of Figure 1.2. Here each coordinate is 0 or 1 depending on whether the corresponding component is functioning properly or has failed. In general, if a system has n components, there are 2^n possible outcomes each of which can be regarded as a point in an n-dimensional sample space. It should be noted that the sample space used

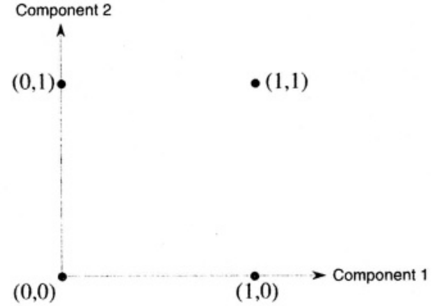


Figure 1.2. A two-dimensional sample space

Figure 1.3. A one-dimensional sample space

here in connection with the observation of the status of components could also serve to describe the results of other experiments; for example, the experiment of observing n successive executions of an **if** statement, with 1 denoting the execution of the **then** clause and 0 denoting the execution of the **else** clause.

The geometric configuration that is used to represent the outcomes of an experiment (e.g., Figure 1.2) is not necessarily unique. For example, we could have regarded the outcomes of the experiment of observing the two-component system to be the total number functioning, and the outcomes would be 0,1,2, as depicted in the one-dimensional sample space of Figure 1.3. Note that point 1 in Figure 1.3 corresponds to points (0,1) and (1,0) in Figure 1.2. It is often easier to use sample spaces whose elements cannot be further "subdivided"; that is, the individual elements of a sample space should not represent two or more outcomes that are distinguishable in some fashion. Thus, sample spaces like those of Figures 1.1 and 1.2 should be used in preference to sample spaces like the one in Figure 1.3.

It is convenient to classify sample spaces according to the number of elements they contain. If the set of all possible outcomes of the experiment is finite, then the associated sample space is a **finite sample space**. Thus, the sample spaces of Figures 1.1–1.3 are finite sample spaces.

To consider an example where a finite sample space does not suffice, suppose we inspect components coming out of an assembly line and that we are interested in the number inspected before we observe the first defective component. It could be the first, the second, ..., the hundredth, ..., and, for all we know, we might have to inspect a billion or more before we find a defective component. Since the number of components to be inspected before the first defective one is found is not known in advance, it is appropriate to take the sample space to be the set of natural numbers. The same sample space results for the experiment of tossing a coin until a head is observed. A sample space such as this, where the set of all outcomes can be put into a one-to-one correspondence with the natural numbers, is said to be **countably infinite**. Usually it is not necessary to distinguish between finite and countably infinite sample spaces. Therefore, if a sample space is either finite or countably infinite, we say that it is a **countable** or a **discrete sample space**.

Measurement of the time until failure of a component would have an entire interval of real numbers as possible values. Since the interval of real numbers cannot be enumerated—that is, they cannot be put into one-to-one correspondence with natural numbers—such a sample space is said to be **uncountable** or **nondenumerable**. If the elements (points) of a sample space constitute a continuum, such as all the points on a line, all the points on a line segment or all the points in a plane, the sample space is said to be **continuous**. Certainly, no real experiment conducted using real measuring devices can ever yield such a continuum of outcomes, since there is a limit to the fineness to which any instrument can measure. However, such a sample space can often be taken as an idealization of, an approximation to, or a model of a real world situation, which may be easier to analyze than a more exact model.

Problems

1. Describe a possible sample space for each of the following experiments:

   (a) A large lot of RAM (random access memory) chips is known to contain a small number of ROM (read-only memory) chips. Three chips are chosen at random from this lot and each is checked to see whether it is a ROM or a RAM.

   (b) A box of 10 chips is known to contain one defective and nine good chips. Three chips are chosen at random from the box and tested.

   (c) An **if**...**then**...**else**... statement is executed 4 times.

1.4 EVENTS

An **event** is simply a collection of certain sample points, that is, a subset of the sample space. Equivalently, any statement of conditions that defines this subset is called an event. Intuitively, an event is defined as a statement whose truth or falsity is determined after the experiment. The set of all experimental outcomes (sample points) for which the statement is true defines the subset of the sample space corresponding to the event. A single performance of the experiment is known as a **trial**. Let E be an event defined on a sample space S; that is, E is a subset of S. Let the outcome of a specific trial be denoted by s, an element of S. If s is an element of E, then we say that the event E has occurred. Only one outcome s in S can occur on any trial. However, every event that includes s will occur.

Consider the experiment of observing a two-component system and the corresponding sample space of Figure 1.2. Let event A_1 be described by the statement "Exactly one component has failed." Then it corresponds to the subset $\{(0,1), (1,0)\}$ of the sample space. We will use the term **event** interchangeably to describe the subset or the statement. There are sixteen different subsets of this sample space with four elements, and each of these subsets defines an event. In particular, the entire sample space $S = \{(0,0),(0,1),(1,0),(1,1)\}$ is an event (called the **universal event**), and so is the null set $\emptyset$ (called the **null** or **impossible event**). The event $\{s\}$ consisting of a single sample point will be called an **elementary event**.

Consider the experiment of observing the time to failure of a component. The sample space, in this case, may be thought of as the set of all nonnegative

TABLE 1.1. Sample Points

$s_0 = (0,0,0,0,0)$	$s_{16} = (1,0,0,0,0)$
$s_1 = (0,0,0,0,1)$	$s_{17} = (1,0,0,0,1)$
$s_2 = (0,0,0,1,0)$	$s_{18} = (1,0,0,1,0)$
$s_3 = (0,0,0,1,1)$	$s_{19} = (1,0,0,1,1)$
$s_4 = (0,0,1,0,0)$	$s_{20} = (1,0,1,0,0)$
$s_5 = (0,0,1,0,1)$	$s_{21} = (1,0,1,0,1)$
$s_6 = (0,0,1,1,0)$	$s_{22} = (1,0,1,1,0)$
$s_7 = (0,0,1,1,1)$	$s_{23} = (1,0,1,1,1)$
$s_8 = (0,1,0,0,0)$	$s_{24} = (1,1,0,0,0)$
$s_9 = (0,1,0,0,1)$	$s_{25} = (1,1,0,0,1)$
$s_{10} = (0,1,0,1,0)$	$s_{26} = (1,1,0,1,0)$
$s_{11} = (0,1,0,1,1)$	$s_{27} = (1,1,0,1,1)$
$s_{12} = (0,1,1,0,0)$	$s_{28} = (1,1,1,0,0)$
$s_{13} = (0,1,1,0,1)$	$s_{29} = (1,1,1,0,1)$
$s_{14} = (0,1,1,1,0)$	$s_{30} = (1,1,1,1,0)$
$s_{15} = (0,1,1,1,1)$	$s_{31} = (1,1,1,1,1)$

real numbers, or the interval $[0, \infty) = \{t \mid 0 \leq t < \infty\}$. Note that this is an example of a continuous sample space. Now if this component is part of a system that is required to carry out a mission of certain duration t, then an event of interest is "The component does not fail before time t." This event may also be denoted by the set $\{x \mid x \geq t\}$, or by the interval $[t, \infty)$.

1.5 ALGEBRA OF EVENTS

Consider an example of a wireless cell with five identical channels. One possible random experiment consists of checking the system to see how many channels are currently available. Each channel is in one of two states: busy (labeled 0) and available (labeled 1). An outcome of the experiment (a point in the sample space) can be denoted by a 5-tuple of 0s and 1s. A 0 in position i of the 5-tuple indicates that channel i is busy and a 1 indicates that it is available. The sample space S has $2^5 = 32$ sample points, as shown in Table 1.1. The event E_1 described by the statement "At least four channels are available" is given by

$$\begin{aligned} E_1 &= \{(0,1,1,1,1),(1,0,1,1,1),(1,1,0,1,1),(1,1,1,0,1),\\ &\quad (1,1,1,1,0),(1,1,1,1,1)\}\\ &= \{s_{15}, s_{23}, s_{27}, s_{29}, s_{30}, s_{31}\}. \end{aligned}$$

The **complement** of this event, denoted by $\overline{E}_1$, is defined to be $S - E_1$, and contains all of the sample points not contained in E_1; that is, $\overline{E}_1 =$

$\{s \in S \mid s \notin E_1\}$. In our example, $\overline{E}_1 = \{s_0$ through s_{14}, s_{16} through s_{22}, s_{24} through $s_{26}, s_{28}\}$. $\overline{E}_1$ may also be described by the statement "at most three channels are available." Let E_2 be the event "at most four channels are available." Then $E_2 = \{s_0$ through $s_{30}\}$. The **intersection** E_3 of the two events E_1 and E_2 is denoted by $E_1 \cap E_2$ and is given by:

$$\begin{aligned} E_3 &= E_1 \cap E_2 \\ &= \{s \in S \mid s \text{ is an element of both } E_1 \text{ and } E_2\} \\ &= \{s \in S \mid s \in E_1 \text{ and } s \in E_2\} \\ &= \{s_{15}, s_{23}, s_{27}, s_{29}, s_{30}\}. \end{aligned}$$

Let E_4 be the event "channel 1 is available." Then $E_4 = \{s_{16}$ through $s_{31}\}$. The **union** E_5 of the two events E_1 and E_4 is denoted by $E_1 \cup E_4$ and is given by:

$$\begin{aligned} E_5 &= E_1 \cup E_4 \\ &= \{s \in S \mid \text{ either } s \in E_1 \text{ or } s \in E_4 \text{ or both}\} \\ &= \{s_{15} \text{ through } s_{31}\}. \end{aligned}$$

Note that E_1 has 6 points, E_4 has 16 points, and E_5 has 17 points. In general:

$$\begin{aligned} |E_5| &= |E_1 \cup E_4| \\ &\le |E_1| + |E_4|. \end{aligned}$$

Here, the notation $|A|$ is used to denote the number of elements in the set A (also known as the **cardinality** of A).

Two events A and B are said to be **mutually exclusive events** or **disjoint events** provided $A \cap B$ is the null set. If A and B are mutually exclusive, then it is not possible for both events to occur on the same trial. For example, let E_6 be the event "channel 1 is busy." Then E_4 and E_6 are mutually exclusive events since $E_4 \cap E_6 = \emptyset$.

Although the definitions of union and intersection are given for two events, we observe that they extend to any finite number of sets. However, it is customary to use a more compact notation. Thus we define

$$\begin{aligned} \bigcup_{i=1}^{n} E_i &= E_1 \cup E_2 \cup E_3 \cdots \cup E_n \\ &= \{s \text{ element of } S | s \text{ element of } E_1 \text{ or } s \text{ element of } E_2 \\ &\quad \text{ or } \cdots s \text{ element of } E_n\} \\ \bigcap_{i=1}^{n} E_i &= E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_n \\ &= \{s \text{ element of } S | s \text{ element of } E_1 \text{ and } s \text{ element of } E_2 \\ &\quad \text{ and } \cdots s \text{ element of } E_n\} \end{aligned}$$

These definitions can also be extended to the union and intersection of a countably infinite number of sets.

The algebra of events may be fully defined by the following five laws or axioms, where A, B, and C are arbitrary sets (or events), and S is the universal set (or event):

(E1) *Commutative laws:*

$$A \cup B = B \cup A, \quad A \cap B = B \cap A.$$

(E2) *Associative laws:*

$$\begin{aligned} A \cup (B \cup C) &= (A \cup B) \cup C, \\ A \cap (B \cap C) &= (A \cap B) \cap C. \end{aligned}$$

(E3) *Distributive laws:*

$$\begin{aligned} A \cup (B \cap C) &= (A \cup B) \cap (A \cup C), \\ A \cap (B \cup C) &= (A \cap B) \cup (A \cap C). \end{aligned}$$

(E4) *Identity laws:*

$$A \cup \emptyset = A, \quad A \cap S = A.$$

(E5) *Complementation laws:*

$$A \cup \overline{A} = S, \quad A \cap \overline{A} = \emptyset.$$

Any relation that is valid in the algebra of events can be proved by using these axioms [(E1–E5)]. Some other useful relations are as follows:

(R1) *Idempotent laws:*

$$A \cup A = A, \quad A \cap A = A.$$

(R2) *Domination laws:*

$$A \cup S = S, \quad A \cap \emptyset = \emptyset.$$

(R3) *Absorption laws:*

$$A \cap (A \cup B) = A, \quad A \cup (A \cap B) = A.$$

(R4) *de Morgan's laws:*

$$\overline{(A \cup B)} = \overline{A} \cap \overline{B}, \quad \overline{(A \cap B)} = \overline{A} \cup \overline{B}.$$

(R5) $$\overline{(\overline{A})} = A.$$

(R6) $$A \cup (\overline{A} \cap B) = A \cup B.$$

From the complementation laws, we note that A and $\overline{A}$ are mutually exclusive since $A \cap \overline{A} = \emptyset$. In addition, A and $\overline{A}$ are collectively exhaustive since any point s (an element of S) is either in $\overline{A}$ or in A. These two notions can be generalized to a list of events.

A list of events $A_1, A_2, \ldots, A_n$ is said to be composed of **mutually exclusive** events if and only if

$$A_i \cap A_j = \begin{cases} A_i & \text{if } i = j, \\ \emptyset & \text{otherwise.} \end{cases}$$

Intuitively, a list of events is composed of mutually exclusive events if there is no point in the sample space that is included in more than one event in the list.

A list of events $A_1, A_2, \ldots, A_n$ is said to be **collectively exhaustive** if and only if

$$A_1 \cup A_2 \cdots \cup A_n = S.$$

Given a list of events that is collectively exhaustive, each point in the sample space is included in at least one event in the list. An arbitrary list of events may be mutually exclusive, collectively exhaustive, both, or neither. For each point s in the sample space S, we may define an event $A_s = \{s\}$. The resulting list of events is mutually exclusive and collectively exhaustive (such a list of events is also called a **partition** of the sample space S). Thus, a sample space may be defined as the mutually exclusive and collectively exhaustive listing of all possible outcomes of an experiment.

Problems

1. Four components are inspected and three events are defined as follows:

   $A =$ "all four components are found defective."

   $B =$ "exactly two components are found to be in proper working order."

   $C =$ "at most three components are found to be defective."

   Interpret the following events:

   (a) $B \cup C$.
   (b) $B \cap C$.
   (c) $A \cup C$.
   (d) $A \cap C$.

2. Use axioms of the algebra of events to prove the relations:

   (a) $A \cup A = A$.
   (b) $A \cup S = S$.
   (c) $A \cap \emptyset = \emptyset$.
   (d) $A \cap (A \cup B) = A$.
   (e) $A \cup (\overline{A} \cap B) = A \cup B$.

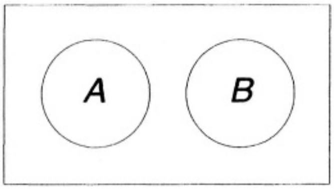


Figure 1.4. Venn diagram for sample space S and events A and B

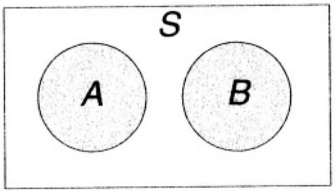


Figure 1.5. Venn diagram of disjoint events A and B

1.6 GRAPHICAL METHODS OF REPRESENTING EVENTS

Venn diagrams often provide a convenient means of ascertaining relations between events of interest. Thus, for a given sample space S and the two events A and B, we have the Venn diagram shown in Figure 1.4. In this figure, the set of all points in the sample space is symbolically denoted by the ones within the rectangle. The events A and B are represented by certain regions in S.

The union of two events A and B is represented by the set of points lying in either A or B. The union of two mutually exclusive events A and B is represented by the shaded region in Figure 1.5. On the other hand, if A and B are not mutually exclusive, they might be represented by a Venn diagram like Figure 1.6. $A \cup B$ is represented by the shaded region; a portion of this shaded region is $A \cap B$ and is so labeled.

For an event A, the complement $\overline{A}$ consists of all points in S that do not belong to A, thus $\overline{A}$ is represented by the unshaded region in Figure 1.7. The usefulness of Venn diagrams becomes apparent when we see that the following

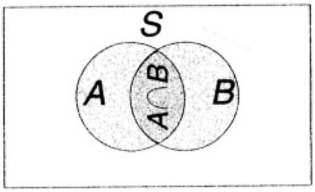


Figure 1.6. Venn diagram for two intersecting events A and B

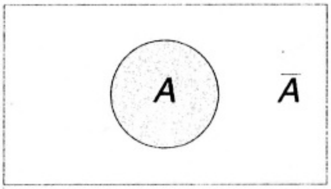


Figure 1.7. Venn diagram of A and its complement

laws of event algebra, discussed in the last section, are easily seen to hold true by reference to Figures 1.6 and 1.7:

$$\begin{aligned} A \cap S &= A, \\ A \cup S &= S, \\ \overline{(\overline{A})} &= A, \\ \overline{(A \cup B)} &= \overline{A} \cap \overline{B}, \\ \overline{(A \cap B)} &= \overline{A} \cup \overline{B}. \end{aligned}$$

Another useful graphical device is the **tree diagram**. As an example, consider the experiment of observing two successive executions of an **if** statement in a certain program. The outcome of the first execution of the **if** statement may be the execution of the **then** clause (denoted by T_1) or the execution of the **else** clause (denoted by E_1). Similarly the outcome of the second execution is T_2 or E_2. This is an example of a **sequential sample space** and leads to the tree diagram of Figure 1.8. We picture the experiment proceeding sequentially downward from the root. The set of all leaves of the tree is the sample space of interest. Each sample point represents the event corresponding to the intersection of all events encountered in tracing a path from the root to the leaf corresponding to the sample point. Note that the four sample points (the leaves of the tree) and their labels constitute the sample space

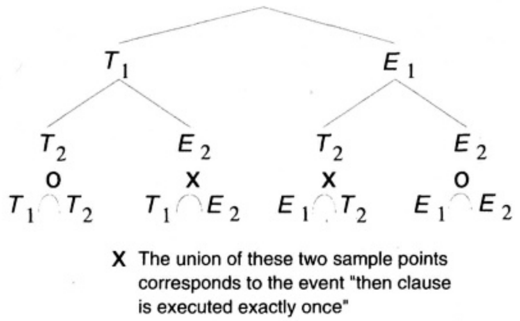

T_1
E_1
T_2
E_2
T_2
E_2
O
X
X
O
$T_1 \cap T_2$
$T_1 \cap E_2$
$E_1 \cap T_2$
$E_1 \cap E_2$
X The union of these two sample points corresponds to the event "then clause is executed exactly once"

Figure 1.8. Tree diagram of a sequential sample space

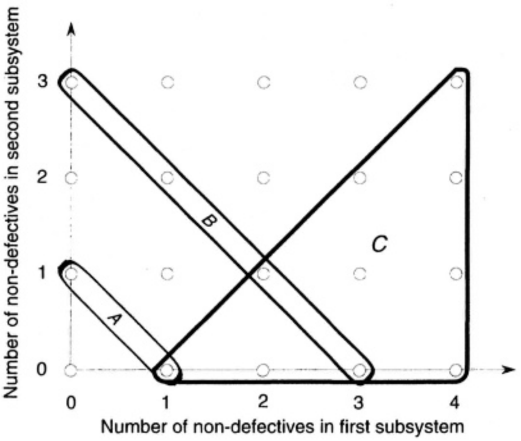

Number of non-defectives in second subsystem
3
2
1
0
0
1
2
3
4
Number of non-defectives in first subsystem
A
B
C

Figure 1.9. A two-dimensional sample space

of the experiment. However, when we deal with a sequential sample space, we normally picture the entire generating tree as well as the resulting sample space.

When the outcomes of the experiment may be expressed numerically, yet another graphical device is a coordinate system. As an example, consider a system consisting of two subsystems. The first subsystem consists of four components and the second subsystem contains three components. Assuming that we are concerned only with the total number of defective components in each subsystem (not with what particular components have failed), the cardinality of the sample space is $5 \cdot 4 = 20$, and the corresponding two-dimensional sample space is illustrated in Figure 1.9. The three events identified in Figure 1.9 are easily seen to be

$A =$"the system has exactly one non-defective component."

$B =$"the system has exactly three non-defective components."

$C =$"the first subsystem has more non-defective components than the second subsystem."

1.7 PROBABILITY AXIOMS

We have seen that the physical behavior of random experiments can be modeled naturally using the concepts of events in a suitably defined sample space. To complete our specification of the model, we shall assign **probabilities** to the events in the sample space. The probability of an event is meant to represent the "relative likelihood" that a performance of the experiment will result in the occurrence of that event. $P(A)$ will denote the probability of the event A in the sample space S.

In many engineering applications and in games of chance, the so-called relative frequency interpretation of the probability is utilized. However, such an approach is inadequate for many applications. We would like the mathematical construction of the probability measure to be independent of the intended application. This leads to an *axiomatic* treatment of the theory of probability. The theory of probability starts with the assumption that probabilities can be assigned so as to satisfy the following three basic **axioms of probability**. The assignment of probabilities is perhaps the most difficult aspect of constructing probabilistic models. Assignments are commonly based on intuition, experience, or experimentation. The theory of probability is neutral; it will make predictions regardless of these assignments. However, the results will be strongly affected by the choice of a particular assignment. Therefore if the assignments are inaccurate, the predictions of the model will be misleading and will not reflect the behavior of the "real world" problem being modeled.

Let S be a sample space of a random experiment. We use the notation $P(A)$ for the probability measure associated with event A. If the event A consists of a single sample point s then $P(A) = P(\{s\})$ will be written as $P(s)$. The probability function $P(\cdot)$ must satisfy the following Kolmogorov's axioms:

(A1) For any event A, $P(A) \geq 0$.

(A2) $P(S) = 1$.

(A3) $P(A \cup B) = P(A) + P(B)$ provided A and B are mutually exclusive events (i.e., when $A \cap B = \emptyset$).

The first axiom states that all probabilities are nonnegative real numbers. The second axiom attributes a probability of unity to the universal event S, thus providing a normalization of the probability measure (the probability of a certain event, an event that must happen, is equal to 1). The third axiom states that the probability function must be additive. These three axioms are easily seen to be consistent with our intuitive ideas of how probabilities behave.

The **principle of mathematical induction** can be used to show [using axiom (A3) as the basis of induction] that for any positive integer n the probability of the union of n mutually exclusive events $A_1, A_2, \ldots, A_n$ is equal to the sum of their probabilities:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i).$$

The three axioms, (A1)–(A3), are adequate if the sample space is finite but to deal with problems with infinite sample spaces, we need to modify axiom A3:

(A3′) For any countable sequence of events $A_1, A_2, \ldots, A_n, \ldots$, that are mutually exclusive (that is, $A_j \cap A_k = \emptyset$ whenever $j \neq k$):

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

All of conventional probability theory follows from the three axioms (A1) through (A3′) of probability measure and the 5 axioms (E1)–(E5) of the algebra of events discussed earlier. These eight axioms can be used to show several useful relations:

(Ra) For any event A, $P(\overline{A}) = 1 - P(A)$.

Proof: A and $\overline{A}$ are mutually exclusive, and $S = A \cup \overline{A}$. Then by axioms (A2) and (A3), $1 = P(S) = P(A) + P(\overline{A})$, from which the assertion follows.

(Rb) If $\emptyset$ is the impossible event, then $P(\emptyset) = 0$.

Proof: Observe that $\emptyset = \overline{S}$ so that the result follows from relation (Ra) and axiom (A2).

(Rc) If A and B are any events, not necessarily mutually exclusive, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: From the Venn diagram of Figure 1.6, we note that $A \cup B = A \cup (\overline{A} \cap B)$ and $B = (A \cap B) \cup (\overline{A} \cap B)$, where the events on the right-hand side are mutually exclusive in each equation. By axiom (A3), we obtain

$$\begin{aligned} P(A \cup B) &= P(A) + P(\overline{A} \cap B) \\ P(B) &= P(A \cap B) + P(\overline{A} \cap B). \end{aligned}$$

The second equation implies $P(\overline{A} \cap B) = P(B) - P(A \cap B)$, which, after substitution in the first equation, yields the desired assertion.

The relation (Rc) can be generalized to a formula similar to the principle of inclusion and exclusion of combinatorial mathematics [LIU 1968]:

(Rd) If $A_1, A_2, \ldots A_n$ are any events, then

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= P(A_1 \cup A_2 \cup \cdots \cup A_n) \\ &= \sum_i P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) \\ &\quad + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) + \cdots \\ &\quad +(-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n), \end{aligned}$$

where the successive sums are over all possible events, pairs of events, triples of events, and so on.

Proof: We prove this result by induction on the number of events n. The result (Rc) above can serve as the basis of induction. Assume inductively that (Rd) holds for a union of $n-1$ events. Define the event $B = A_1 \cup A_2 \cup \cdots \cup A_{n-1}$. Then

$$\bigcup_{i=1}^{n} A_i = B \cup A_n.$$

Using the result (Rc) above, we get

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= P(B \cup A_n) \\ &= P(B) + P(A_n) - P(B \cap A_n). \end{aligned} \tag{1.1}$$

Now

$$B \cap A_n = (A_1 \cap A_n) \cup (A_2 \cap A_n) \cup \cdots \cup (A_{n-1} \cap A_n)$$

is a union of $n-1$ events and hence, using the inductive hypothesis, we get

$$\begin{aligned} P(B \cap A_n) &= P(A_1 \cap A_n) + P(A_2 \cap A_n) + \cdots + P(A_{n-1} \cap A_n) \\ &\quad -P[(A_1 \cap A_n) \cap (A_2 \cap A_n)] \\ &\quad -P[(A_1 \cap A_n) \cap (A_3 \cap A_n)] \\ &\quad - \cdots \\ &\quad +P[(A_1 \cap A_n) \cap (A_2 \cap A_n) \cap (A_3 \cap A_n)] \\ &\quad + \cdots - \cdots \\ &\quad +(-1)^{n-2} P[(A_1 \cap A_n) \cap (A_2 \cap A_n) \cap \cdots \cap (A_{n-1} \cap A_n)] \\ &= P(A_1 \cap A_n) + P(A_2 \cap A_n) + \cdots + P(A_{n-1} \cap A_n) \\ &\quad -P(A_1 \cap A_2 \cap A_n) - P(A_1 \cap A_3 \cap A_n) - \cdots \\ &\quad +P(A_1 \cap A_2 \cap A_3 \cap A_n) + \cdots \\ &\quad - \cdots \\ &\quad +(-1)^{n-2} P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_{n-1} \cap A_n). \end{aligned} \tag{1.2}$$

Also, since $B = A_1 \cup A_2 \cup \cdots \cup A_{n-1}$ is a union of $n-1$ events, the inductive hypothesis gives

$$\begin{aligned} P(B) &= P(A_1) + P(A_2) + \cdots + P(A_{n-1}) \\ &\quad -P(A_1 \cap A_2) - P(A_1 \cap A_3) - \cdots \\ &\quad + \cdots \\ &\quad +(-1)^{n-2} P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \end{aligned} \tag{1.3}$$

Substituting (1.2) and (1.3) into (1.1), we obtain the required result.

The relation (Rd) is computationally expensive to use. A computationally simpler formula is the sum of disjoint products (SDP) formula below.

(Re)

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= P(A_1) + P(\overline{A}_1 \cap A_2) + P(\overline{A}_1 \cap \overline{A}_2 \cap A_3) + \cdots \\ &\quad +P(\overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_{n-1} \cap A_n). \end{aligned} \tag{1.4}$$

The SDP formula is frequently used in reliability computations [LUO 1998]. We leave the proof as an exercise.

To avoid certain mathematical difficulties, we must place restrictions on which subsets of the sample space may be termed *events* to which probabilities can be assigned. In a given problem there will be a particular class of subsets of S that is "measurable" and will be called the "class of events" $\mathcal{F}$. Since we would like to perform the standard set operations on events, it is reasonable to demand that $\mathcal{F}$ be closed under countable unions of events in $\mathcal{F}$ as well as under complementation. A collection of subsets of a given set S that is closed under countable unions and complementation is called a σ field of subsets of S. Now a **probability space** or **probability system** may be defined as a triple $(S, \mathcal{F}, P)$, where S is a set, $\mathcal{F}$ is a σ-field of subsets of S, and P is a probability measure on $\mathcal{F}$ assumed to satisfy axioms (A1)–(A3$'$).

If the sample space is discrete (finite or countable), then every subset of S can be an event belonging to $\mathcal{F}$. However, in the case that S is uncountable, this is no longer true. For example, let S be the interval [0,1] and assume the probability assignment $P(a \leq x \leq b) = b - a$ for $0 \leq a \leq b \leq 1$. Then it can be shown that not all possible subsets of S can be assigned a probability in a manner consistent with the three axioms of P. In such cases, the smallest σ field of subsets of S containing all open and closed intervals is usually adopted as the class of events $\mathcal{F}$.

In summary, P is a function with domain $\mathcal{F}$ and range $[0, 1]$, which satisfies the three axioms (A1), (A2), and (A3$'$). P assigns a number between 0 and 1 to any event in $\mathcal{F}$. In general, $\mathcal{F}$ does not include all possible subsets of S, and the subsets (events) included in $\mathcal{F}$ are called *measurable*. However, for our purposes, every subset of a sample space constructed here can be considered an event having a probability.

We now outline the steps of a basic procedure to be followed in solving problems [GOOD 1977]:

1. *Identify the sample space S.* The sample space S must be chosen so that all of its elements are mutually exclusive and collectively exhaustive, that is, no two elements can occur simultaneously and one element must occur on any trial. Many of the "trick" probability problems are based on some ambiguity in the problem statement or an inexact formulation of the model of a physical situation. The choice of an appropriate sample space resulting from a detailed description of the model, will do much to resolve common difficulties. Since many choices for the sample space are possible, it is advisable to use a sample space whose elements cannot be further "subdivided" — that is, all possible distinguishable outcomes of the experiment should be listed separately.

2. *Assign probabilities.* Assign probabilities to the elements in S. This assignment must be consistent with the axioms (A1), (A2), and (A3). In practice, probabilities are assigned either on the basis of estimates obtained from past experience, or on the basis of a careful analysis of

conditions underlying the random experiment, or on the basis of assumptions, such as the common assumption that various outcomes in a finite sample space are equiprobable (equally likely).

3. *Identify the events of interest.* The events of interest, in a practical situation, will be described by statements. These need to be recast as subsets of the sample space. The laws of event algebra (E1)–(E5) and (R1)–(R6) may be used for any simplification. Pictorial devices such as Venn diagrams, tree diagrams, or coordinate system plots may also be used to advantage.

4. *Compute desired probabilities.* Calculate the probabilities of the events of interest using the axioms (A1), (A2), and (A3′) and any derived laws such as (Ra), (Rb), (Rc), (Rd), and (Re). It is usually helpful to express the event of interest as a union of mutually exclusive points in the sample space and summing the probabilities of all points included in the union.

Example 1.1

As a simple illustration of this procedure, consider the example of the wireless cell with 5 channels.

Step 1: An appropriate sample space consists of 32 points (see Table 1.1), each represented by a 5-tuple of 0s and 1s. A 0 in position i indicates that channel i is busy and a 1 indicates that it is available.

Step 2: In the absence of detailed knowledge about the system, we assume that each sample point is equally likely. Since there are 32 sample points, we assign a probability of $\frac{1}{32}$ to each, that is, $P(s_0) = P(s_1) = \cdots = P(s_{31}) = \frac{1}{32}$. It is easily seen that this assignment is consistent with the three probability axioms.

Step 3: Assume that we are required to determine the probability that a call is not blocked, given that the conference call needs at least three channels for its execution. The event E of interest, then, is "three or more channels are available." From the definition of the sample points, we see that

$$\begin{aligned} E &= \{s_7, s_{11}, s_{13}, s_{14}, s_{15}, s_{19}, s_{21}, s_{22}, s_{23}, s_{25} - -s_{31}\} \\ &= \{s_7\} \cup \{s_{11}\} \cup \{s_{13}\} \cup \{s_{14}\} \cup \{s_{15}\} \cup \{s_{19}\} \cup \{s_{21}\} \cup \{s_{22}\} \\ &\quad \cup \{s_{23}\} \cup \{s_{25}\} \cup \{s_{26}\} \cup \{s_{27}\} \cup \{s_{28}\} \cup \{s_{29}\} \cup \{s_{30}\} \\ &\quad \cup \{s_{31}\}. \end{aligned}$$

Step 4: We have already simplified E so that it is expressed as a union of mutually exclusive events. The probability of each of these elementary events is $\frac{1}{32}$. Thus, a repeated application of axiom (A3′) gives us

$$\begin{aligned} P(E) &= \sum_{s_i \in E} P(s_i) \\ &= \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} \\ &\quad + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} + \frac{1}{32} \\ &= \frac{1}{2}. \end{aligned}$$

Alternatively, we could have noted that E consists of 16 sample points and since each 32 sample point is equally likely, $P(E) = \frac{16}{32}$.

♯

Problems

1. Give the proof of the relation (Re) in this section.
2. Consider a pool of six I/O (input/output) buffers. Assume that any buffer is just as likely to be available (or occupied) as any other. Compute the probabilities associated with the following events:

   $A =$ "at least 2 but no more than 5 buffers occupied."

   $B =$ "at least 3 but no more than 5 occupied."

   $C =$ "all buffers available or an even number of buffers occupied."

   Also determine the probability that at least one of the events A, B, and C occurs.
3. Show that if event B is contained in event A, then $P(B) \leq P(A)$.

1.8 COMBINATORIAL PROBLEMS

If the sample space of an experiment consists of only a finite number n of sample points, or **elementary events**, then the computation of probabilities is often simple. Assume that assignment of probabilities is made such that for s_i (an element of S), $P(s_i) = p_i$ and

$$\sum_{i=1}^{n} p_i = 1.$$

Since any event E consists of a certain collection of these sample points, $P(E)$ can be found, using axiom (A3′), by adding up the probabilities of the separate sample points that make up E (recall the wireless cell example of the last section).

Example 1.2

Consider the following **if** statement in a program:

if B **then** s_1 **else** s_2.

The random experiment consists of "observing" two successive executions of the **if** statement. The sample space consists of the four possible outcomes:

$$\begin{aligned} S &= \{(s_1, s_1), (s_1, s_2), (s_2, s_1), (s_2, s_2)\} \\ &= \{t_1, t_2, t_3, t_4\}. \end{aligned}$$

Assume that on the basis of strong experimental evidence, the following probability assignment is justified:

$$P(t_1) = 0.34, P(t_2) = 0.26, P(t_3) = 0.26, P(t_4) = 0.14.$$

The events of interest are given E_1 = "at least one execution of the statement s_1" and E_2 = "statement s_2 is executed the first time." It is easy to see that

$$\begin{aligned} E_1 &= \{(s_1, s_1), (s_1, s_2), (s_2, s_1)\} \\ &= \{t_1, t_2, t_3\}, \\ E_2 &= \{(s_2, s_1), (s_2, s_2)\} \\ &= \{t_3, t_4\}, \\ P(E_1) &= P(t_1) + P(t_2) + P(t_3) = 0.86, \\ P(E_2) &= P(t_3) + P(t_4) = 0.4. \end{aligned}$$

♯

In the special case when $S = \{s_1, \ldots s_n\}$ and $P(s_i) = p_i = (1/n)$ (equally likely sample points), the situation is even simpler. Calculation of probabilities is then reduced to simply counting the number of sample points in the event of interest. If the event E consists of k sample points, then

$$\begin{aligned} P(E) &= \frac{\text{number of points in } E}{\text{number of points in } S} \\ &= \frac{\text{favorable outcomes}}{\text{total outcomes}} \\ &= \frac{k}{n}. \end{aligned} \tag{1.5}$$

Example 1.3

A group of four VLSI chips consists of two good chips, labeled g_1 and g_2, and two defective chips, labeled d_1 and d_2. If three chips are selected at random from this group, what is the probability of the event E = "two of the three selected chips are defective"?

A natural sample space for this problem consists of all possible three chip selections from the group of four chips: $S = \{g_1g_2d_1, g_1g_2d_2, g_1d_1d_2, g_2d_1d_2\}$. It is customary to interpret the phrase "selected at random" as implying equiprobable sample points. Since the two sample points $g_1d_1d_2$ and $g_2d_1d_2$ are favorable to the event E, and since the sample space has four points, we conclude that $P(E) = \frac{2}{4} = \frac{1}{2}$.

♯

We have seen that under the equiprobability assumption, finding $P(E)$ simply involves counting the number of outcomes favorable to E. However, counting by hand may not be feasible when the sample space is large. Standard counting methods of combinatorial analysis can often be used to avoid writing down the list of favorable outcomes explicitly.

1.8.1 Ordered Samples of Size k, with Replacement

Here we are interested in counting the number of ways we can select k objects from among n objects where order is important and when the same object is allowed to be repeated any number of times **(permutations with replacement)**. Alternatively, we are interested in the number of ordered sequences $(s_{i_1}, s_{i_2}, \ldots, s_{i_k})$, where each s_{i_r} belongs to $\{s_1, \ldots, s_n\}$. It is not difficult to see that the required number is $\big(n \cdot n \cdot \cdots \cdot n(k\text{times})\big)$, or n^k.

Example 1.4

Assume that we are interested in finding the probability that some randomly chosen k-digit decimal number is a valid k-digit octal number. The sample space, in this case, is

$$S = \{(x_1, x_2, \ldots, x_k) \mid x_i \in \{0, 1, 2, \ldots, 9\}\}$$

and the event of interest is

$$E = \{(x_1, x_2 \ldots x_k) \mid x_i \in \{0, 1, 2, \ldots, 7\}\}.$$

By the above counting principle, $|S| = 10^k$ and $|E| = 8^k$. Now, if we assume that all the sample points are equally likely, then the required answer is

$$P(E) = \frac{|E|}{|S|} = \frac{8^k}{10^k} = \frac{4^k}{5^k}.$$

♯

1.8.2 Ordered Samples of Size k, without Replacement

The number of ordered sequences $(s_{i_1}, s_{i_2}, \ldots, s_{i_k})$, where each s_{i_r} belongs to $\{s_1, \ldots, s_n\}$, but repetition is not allowed (i.e., no s_i can appear more than once in the sequence), is given by

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!} \qquad \text{for } k = 1, 2, \ldots, n.$$

This number is also known as the number of permutations of n distinct objects taken k at a time, and denoted by $P(n, k)$.

Example 1.5

Suppose we wish to find the probability that a randomly chosen three-letter sequence will not have any repeated letters.

Let $I = \{a, b, \ldots, z\}$ be the alphabet of 26 letters. Then the sample space is given by

$$S = \{(\alpha, \beta, \gamma) \mid \alpha \in I, \beta \in I, \gamma \in I\}$$

and the event of interest is

$$E = \{(\alpha, \beta, \gamma) \mid \alpha \in I, \beta \in I, \gamma \in I, \alpha \neq \beta, \beta \neq \gamma, \alpha \neq \gamma\}.$$

By the abovementioned counting principle, $|E|$ is simply $P(26, 3) = 15,600$. Furthermore, $|S| = 26^3 = 17,576$. Therefore, the required answer is

$$P(E) = \frac{15,600}{17,576} = 0.8875739.$$

♯

1.8.3 Unordered Samples of Size k, without Replacement

The number of unordered sets $\{s_{i_1}, s_{i_2}, \ldots, s_{i_k}\}$, where s_{i_r} $(r = 1, 2, \ldots, k)$ are distinct elements of $\{s_1, \ldots, s_n\}$ is

$$\frac{n!}{k!(n-k)!} \quad \text{for } k = 0, 1, \ldots, n.$$

This is also known as the number of combinations of n distinct objects taken k at a time, and is denoted by $\binom{n}{k}$.

Example 1.6

If a box contains 75 good VLSI chips and 25 defective chips, and 12 chips are selected at random, find the probability that at least one chip is defective.

By the counting principle described above, the number of unordered samples without replacement is $\binom{100}{12}$ and hence the size of the sample space is $|S| = \binom{100}{12}$. The event of interest is E = "at least one chip is defective." Here we find it easier to work with the complementary event $\overline{E}$ = "no chip is defective." Since there are 75 good chips, the preceding counting principle yields $|\overline{E}| = \binom{75}{12}$. Then

$$\begin{aligned} P(\overline{E}) &= \frac{|\overline{E}|}{|S|} \\ &= \frac{\binom{75}{12}}{\binom{100}{12}} \\ &= \frac{75! \cdot 12! \cdot 88!}{12! \cdot 63! \cdot 100!} \\ &= \frac{75! \cdot 88!}{63! \cdot 100!}. \end{aligned}$$

Now since $P(E) = 1 - P(\overline{E})$, the required probability is easily obtained.

♯

Example 1.7

Consider a TDMA (time division multiple access) wireless system [SUN 1999], where the base transceiver system of each cell has n base repeaters [also called base radio (BR)]. Each base repeater provides m time-division-multiplexed channels. Thus, there are mn channels in the system. We note that normally a cell reserves one or more channels for signaling transfer, which resides in one of n base repeaters. However, for simplicity, we do not consider signaling channels (also called *control channels*) in this example.

A base repeater is subject to failure. In order to evaluate the impact of such a failure on the performability of the system, we should know the number of ongoing talking channels on the failed base repeater. Suppose the channels are allocated randomly to the users. Denote the total number of talking channels in the whole system as k, and the number of idle channels in the whole system as j ($j + k = mn$ always holds), when the failure occurs. Then the probability, p_i, that i talking channels reside in the failed base repeater is given by

$$p_i = \frac{\binom{m(n-1)}{k-i}\binom{m}{i}}{\binom{mn}{k}}, \quad \text{for } 0 \leq i \leq \min(m, k). \tag{1.6}$$

Clearly, the total number of possible combinations to have k talking channels out of mn channels is $\binom{mn}{k}$, namely, the size of the sample space, $|S|$. The event of interest is E = "i talking channels on the failed base repeater." Now if i $(0 \leq i \leq \min(m, k))$ out of the k talking channels are on the failed base repeater, corresponding to a total of $\binom{m}{i}$ combinations, then $(k - i)$ talking channels are on the rest of the $(n - 1)$ base repeaters, which has $\binom{m(n-1)}{k-i}$ combinations. Thus, $|E| = \binom{m(n-1)}{k-i}\binom{m}{i}$. Probability p_i can now be easily obtained as $|E|/|S|$.

♯

Problems

1. How many even two-digit numbers can be constructed out of the digits 3, 4, 5, 6, and 7? Assume first that you may use the same digit again. Next, answer this question assuming that you cannot use a digit more than once.

2. Three couples (husbands and their wives) must sit at a round table in such a way that no husband is placed next to his wife. How many configurations exist?. If seats are occupied at random, what is the probability of such a configuration?

3. If a three-digit decimal number is chosen at random, find the probability that exactly k digits are ≥ 5, for $0 \leq k \leq 3$.

4. A box with 15 VLSI chips contains five defective ones. If a random sample of three chips is drawn, what is the probability that all three are defective?

5. In a party of five persons, compute the probability that at least two of the persons have the same birthday (month/day), assuming a 365-day year.

6. * A series of n jobs arrive at a multiprocessor computer with n processors. Assume that each of the n^n possible assignment vectors (processor for job 1, ..., processor for job n) is equally likely. Find the probability that exactly one processor will not be assigned a job.

1.9 CONDITIONAL PROBABILITY

So far, we have assumed that the only information about the outcome of a trial of a given experiment, available before the trial, is that the outcome will correspond to some point in the sample space S. With this assumption, we can compute the probability of some event A. Suppose that we are given the added information that the outcome s of a trial is contained in a subset B of the sample space, with $P(B) \neq 0$. Knowledge of the occurrence of the event B may change the probability of the occurrence of the event A. We wish to define the **conditional probability** of the event A given that the event B occurs, or the **conditional probability of A given B**, symbolically as $P(A|B)$. Given that event B has occurred, the sample point corresponding to the outcome of the trial must be in B and cannot be in $\overline{B}$. To reflect this partial information, we define the conditional probability of a sample point s (an element of S) by

$$P(s|B) = \begin{cases} \frac{P(s)}{P(B)} & \text{if } s \in B, \\ 0 & \text{if } s \in \overline{B}. \end{cases}$$

Thus the original probability assigned to a sample point in B is scaled up by $1/P(B)$, so that the probabilities of the sample points in B will add up to 1. Now the conditional probability of any other event, such as A, can be obtained by summing over the conditional probabilities of the sample points included in A (noting that $A = (A \cap B) \cup (A \cap \overline{B})$):

$$\begin{aligned} P(A|B) &= \sum_{s \in A} P(s|B) \\ &= \sum_{s \in A \cap \overline{B}} P(s|B) + \sum_{s \in A \cap B} P(s|B) \\ &= \sum_{s \in A \cap B} P(s|B) \\ &= \sum_{s \in A \cap B} \frac{P(s)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0. \end{aligned}$$

This leads us to the following definition.

Definition (Conditional Probability). The conditional probability of A given B is

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

if $P(B) \neq 0$ and it is undefined otherwise.

A rearrangement of this definition yields the following **multiplication rule (MR)**:

$$P(A \cap B) = \begin{cases} P(B)P(A|B) & \text{if } P(B) \neq 0, \\ P(A)P(B|A) & \text{if } P(A) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Example 1.8

We are given a box containing 5000 VLSI chips, 1000 of which are manufactured by company X and the rest by company Y. Ten percent of the chips made by company X are defective and 5% of the chips made by company Y are defective. If a randomly chosen chip is found to be defective, find the probability that it came from company X.

Define the events $A =$ "chip made by company X" and $B =$ "chip is defective." Since out of 5000 chips, 1000 are made by company X, we conclude that $P(A) = 1000/5000 = 0.2$. Also, out of a total of 5000 chips, 300 are defective. Therefore, $P(B) = 300/5000 = 0.06$. Now the event $A \cap B =$ "the chip is made by company X and is defective." Out of 5000 chips, 100 chips qualify for this statement. Thus $P(A \cap B) = 100/5000 = 0.02$. Now the quantity of interest is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.02}{0.06} = \frac{1}{3}.$$

Thus the knowledge of the occurrence of B has increased the probability of the occurrence of event A. Similarly we find that the knowledge of the occurrence of A has increased the chances for the occurrence of the event B, since $P(B|A) = 0.1$. In fact, note that

$$\frac{P(A|B)}{P(B|A)} = \frac{1/3}{0.1} = \frac{0.2}{0.06} = \frac{P(A)}{P(B)}.$$

This interesting property of conditional properties is easily shown to hold in general

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

♯

Problems

1. Consider four computer firms, A, B, C, D, bidding for a certain contract. A survey of past bidding success of these firms on similar contracts shows the following probabilities of winning:

$$P(A) = 0.35, P(B) = 0.15, P(C) = 0.3, P(D) = 0.2.$$

Before the decision is made to award the contract, firm B withdraws its bid. Find the new probabilities of A, C, D winning the bid.

1.10 INDEPENDENCE OF EVENTS

We have seen that it is possible for the probability of an event A to decrease or increase given that event B has occurred. If the probability of the occurrence of an event A does not change regardless of whether event B has occurred, we are likely to conclude that the two events are independent. Thus we define two events A and B to be independent if and only if

$$P(A|B) = P(A).$$

From the definition of conditional probability, we have [provided $P(A) \neq 0$ and $P(B) \neq 0$]:

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

From this we conclude that the condition for the independence of A and B can also be given either as $P(A|B) = P(A)$ or as $P(A \cap B) = P(A)P(B)$. Note that $P(A \cap B) = P(A)P(B|A)$ (if $P(A) \neq 0$) holds regardless of whether A and B are independent, but $P(A \cap B) = P(A)P(B)$ holds only when A and B are independent. In fact this latter condition is the usual definition of independence.

Definition (Independent Events). Events A and B are said to be independent if

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

This equation is symmetric in A and B and shows that whenever A is independent of B, so is B of A. Some authors use the phrases "stochastically independent events" or "statistically independent events" in place of just "independent events." Note that if A and B are not independent, then $P(A \cap B)$ is computed using the multiplication rule of the last section. The abovementioned condition for independence can be derived in another way by first noting that the event A is a disjoint union of events $A \cap B$ and $A \cap \overline{B}$. Now the conditional probability of all the sample points in the latter event is zero while the conditional probability of all the sample points in the former event is increased by the factor $1/P(B)$. Therefore, for $P(A|B) = P(A)$ to hold, the decrease in probability due to points in $A \cap \overline{B}$ must be balanced by the increase in probability due to points in $A \cap B$. In other words

$$\frac{P(A \cap B)}{P(B)} - P(A \cap B) = P(A \cap \overline{B}) - 0$$

or

$$\begin{aligned}\frac{P(A \cap B)}{P(B)} &= P(A \cap \overline{B}) + P(A \cap B)\\ &= P(A).\end{aligned}$$

Example 1.9

A microcomputer system consists of a microprocessor CPU chip and a random access main memory chip. The CPU is selected from a lot of 100, 10 of which are defective, and the memory chip is selected from a lot of 300, 15 of which are defective. Define A to be the event "the selected CPU is defective," and let B be the event "the selected memory chip is defective." Then $P(A) = 10/100 = 0.1$, and $P(B) = 15/300 = 0.05$. Since the two chips are selected from different lots, we may expect the events A and

B to be independent. This can be checked since there are $10 \cdot 15$ ways of choosing both defective chips and there are $100 \cdot 300$ ways of choosing two chips. Thus

$$\begin{aligned} P(A \cap B) &= \frac{10 \cdot 15}{100 \cdot 300} \\ &= 0.005 \\ &= 0.10 \cdot 0.05 \\ &= P(A)P(B). \end{aligned}$$

♯

Several important points are worth noting about the concept of independence:

1. If A and B are two mutually exclusive events, then $A \cap B = \emptyset$, which implies $P(A \cap B) = 0$. Now if they are independent as well, then either $P(A) = 0$ or $P(B) = 0$.
2. If an event A is independent of itself, that is, if A and A are independent, then $P(A) = 0$ or $P(A) = 1$, since the assumption of independence yields $P(A \cap A) = P(A)P(A)$ or $P(A) = [P(A)]^2$.
3. If the events A and B are independent and the events B and C are independent, then events A and C need not be independent. In other words, the relation of independence is not a transitive relation.
4. If the events A and B are independent, then so are events $\overline{A}$ and B, events A and $\overline{B}$, and events $\overline{A}$ and $\overline{B}$. To show the independence of events $\overline{A}$ and B, note that $A \cap B$ and $\overline{A} \cap B$ are mutually exclusive events whose union is B. Therefore

$$\begin{aligned} P(B) &= P(A \cap B) + P(\overline{A} \cap B) \\ &= P(A)P(B) + P(\overline{A} \cap B) \end{aligned}$$

since A and B are independent. This implies that $P(\overline{A} \cap B) = P(B) - P(A)P(B) = P(B)[1 - P(A)] = P(B)P(\overline{A})$, which establishes the independence of $\overline{A}$ and B. Independence of A and $\overline{B}$, and $\overline{A}$ and $\overline{B}$ can be shown similarly.

The concept of independence of two events can be naturally extended to a list of n events.

Definition (Independence of a Set of Events). A list of n events $A_1, A_2, \ldots, A_n$ is defined to be mutually independent if and only if for each set of k $(2 \le k \le n)$ distinct indices $i_1, i_2, \ldots, i_k$, which are elements of $\{1, 2, \ldots, n\}$, we have

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

Given that a list of events $A_1, A_2, \ldots, A_n$ is mutually independent, it is straightforward to show that for each set of distinct indices $i_1, i_2, \ldots, i_k$, which are elements of $\{1, 2, \ldots, n\}$:

$$P(B_{i_1} \cap B_{i_2} \cap \cdots \cap B_{i_k}) = P(B_{i_1})P(B_{i_2}) \cdots P(B_{i_k}) \tag{1.7}$$

where each B_{i_k} may be either A_{i_k} or $\overline{A}_{i_k}$. In other words, if the A_is are independent and we replace any event by its complement, we still have independence.

By the probability axiom (A3), if a list of events is mutually exclusive, the probability of their union is the sum of their probabilities. On the other hand, if a list of events is mutually independent, the probability of their intersection is the product of their probabilities. The additive and multiplicative nature, respectively, of two event lists should be noted.

Note that it is possible to have $P(A \cap B \cap C) = P(A)P(B)P(C)$ with $P(A \cap B) \neq P(A)P(B)$, $P(A \cap C) \neq P(A)P(C)$, and $P(B \cap C) \neq P(B)P(C)$. Under these conditions, events A, B, and C are not mutually independent. Similarly, the condition $P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n)$ does not imply a similar condition for any smaller family of events, and therefore this condition does not imply that events $A_1, A_2, \ldots, A_n$ are mutually independent.

Example 1.10 [ASH 1970]

Consider the experiment of rolling two dice. Let the sample space $S = \{(i, j) \mid 1 \leq i, j \leq 6\}$. Also assume that each sample point is assigned a probability of $\frac{1}{36}$. Define the events A, B, and C so that

$A =$ "first die results in a 1, 2, or 3."

$B =$ "first die results in a 3, 4, or 5."

$C =$ "the sum of the two faces is 9."

Then $A \cap B = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\}$, $A \cap C = \{(3,6)\}$, $B \cap C = \{(3,6), (4,5), (5,4)\}$, and $A \cap B \cap C = \{(3,6)\}$. Therefore

$$\begin{aligned} P(A \cap B) &= \frac{1}{6} \neq P(A)P(B) = \frac{1}{4}, \\ P(A \cap C) &= \frac{1}{36} \neq P(A)P(C) = \frac{1}{18}, \\ P(B \cap C) &= \frac{1}{12} \neq P(B)P(C) = \frac{1}{18}, \end{aligned}$$

but

$$P(A \cap B \cap C) = \frac{1}{36} = P(A)P(B)P(C).$$

♯

If the events $A_1, A_2, \ldots, A_n$ are such that every pair is independent, then such events are called **pairwise independent**. It does not follow that the list of events is **mutually independent**.

Example 1.11 [ASH 1970]

Consider the above experiment of tossing two dice. Let

$A =$ "first die results in a 1, 2, or 3."

$B =$ "second die results in a 4, 5, or 6."

$C =$ "the sum of the two faces is 7."

Then

$$A \cap B = \{(1,4),(1,5),(1,6),(2,4),(2,5),(2,6),(3,4),(3,5),(3,6)\}$$

and

$$\begin{aligned} A \cap C &= B \cap C \\ &= A \cap B \cap C \\ &= \{(1,6),(2,5),(3,4)\}. \end{aligned}$$

Therefore

$$\begin{aligned} P(A \cap B) &= \frac{1}{4} = P(A)P(B), \\ P(A \cap C) &= \frac{1}{12} = P(A)P(C), \\ P(B \cap C) &= \frac{1}{12} = P(B)P(C), \end{aligned}$$

but

$$P(A \cap B \cap C) = \frac{1}{12} \neq P(A)P(B)P(C) = \frac{1}{24}.$$

In this example, events A, B, and C are **pairwise independent** but not **mutually independent.**

♯

We illustrate the idea of independence by considering the problem of computing reliability of so-called series–parallel systems. A **series system** is one in which all components are so interrelated that the entire system will fail if any one of its components fails. On the other hand, a **parallel system** is one that will fail only if all of its components fail. We will assume that failure events of components in a system are mutually independent.

First consider a series system of n components. For $i = 1, 2, \ldots, n$, define events $A_i =$ "component i is functioning properly." Let the **reliability**, R_i, of component i be defined as the probability that the component is functioning properly; then $R_i = P(A_i)$. By the assumption of series connections, the system reliability:

$$\begin{aligned} R_s &= P(\text{"the system is functioning properly"}) \\ &= P(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &= P(A_1)P(A_2)\cdots P(A_n) \\ &= \prod_{i=1}^{n} R_i. \end{aligned} \tag{1.8}$$

This simple **product law of reliabilities,** applicable to series systems of independent components, demonstrates how quickly system reliability degrades with an increase in complexity. For example, if a system consists of five components each in series, each having a reliability of 0.970, then the system reliability is $0.970^5 = 0.859$. Now if the system complexity is increased so that it contained 10 similar components, its reliability would be reduced to $0.970^{10} = 0.738$. Consider what happens to system reliability when a large system like a computer system consists of tens to hundreds of thousands of components!

One way to increase the reliability of a system is to use **redundancy**. The first scheme that comes to mind is to replicate components with small reliabilities (**parallel redundancy**). First consider a system consisting of n independent components in parallel, so that it will fail to function only if all n components have failed. Define event $A_i =$ "the component i is functioning properly" and $A_p =$ "the parallel system of n components is functioning properly." Also let $R_i = P(A_i)$ and $R_p = P(A_p)$. To establish a relation between A_p and the A_i values, it is easier to consider the complementary events. Thus

$$\begin{aligned}\overline{A}_p &= \text{"the parallel system has failed"}\\ &= \text{"all } n \text{ components have failed"}\\ &= \overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n.\end{aligned}$$

Therefore

$$\begin{aligned}P(\overline{A}_p) &= P(\overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n)\\ &= P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n)\end{aligned}$$

by independence. Now let $F_p = 1 - R_p$ be the **unreliability** of the parallel system and similarly let $F_i = 1 - R_i$ be the unreliability of component i. Then, since A_i and $\overline{A}_i$ are mutually exclusive and collectively exhaustive events, we have

$$\begin{aligned}1 &= P(S)\\ &= P(A_i) + P(\overline{A}_i)\end{aligned}$$

and

$$\begin{aligned}F_i &= P(\overline{A}_i)\\ &= 1 - P(A_i).\end{aligned}$$

Then

$$\begin{aligned}F_p &= P(\overline{A}_p)\\ &= \prod_{i=1}^{n} F_i\end{aligned}$$

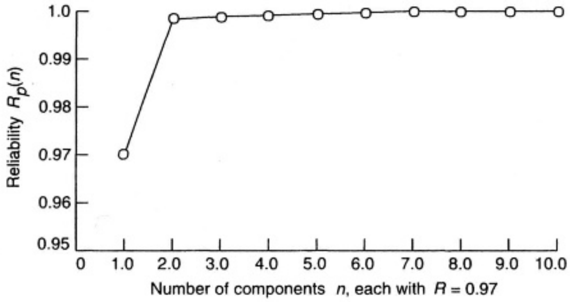


Figure 1.10. Reliability curve of a parallel redundant system

and

$$\begin{aligned} R_p &= 1 - F_p \\ &= 1 - \prod_{i=1}^{n}(1 - R_i). \end{aligned} \tag{1.9}$$

Thus, for parallel systems of n independent components, we have a **product law of unreliabilities** analogous to the product law of reliabilities of series systems. If we have a parallel system of five components, each with a reliability of 0.970, then the system reliability is increased to

$$\begin{aligned} 1 - (1 - 0.970)^5 &= 1 - (0.03)^5 \\ &= 1 - 0.0000000243 \\ &= 0.9999999757. \end{aligned}$$

However, one should be aware of a **law of diminishing returns**, according to which the rate of increase in reliability with each additional component decreases rapidly as n increases. This is illustrated in Figure 1.10, where we have plotted R_p as a function of n. [This remark is easily formalized by noting that R_p is a concave function of n since $R_p'(n) = -(1 - R)^n \ln(1 - R) > 0$, and $R_p''(n) = -(1 - R)^n \big(\ln(1 - R)\big)^2 < 0$.]

The basic formulas (1.8) and (1.9) for the reliability computation of series and parallel systems can be used in combination to compute the reliability of a system having both series and parallel parts (**series–parallel systems**). Consider a series–parallel system of n serial stages where stage i consists of n_i identical components in parallel. Let the reliability of each component at stage i be R_i. Assuming that all components are independent, system reliability R_{sp} can be computed from the formula

$$R_{sp} = \prod_{i=1}^{n}\left[1 - (1 - R_i)^{n_i}\right]. \tag{1.10}$$

A series–parallel system can be graphically represented by a series–parallel reliability block diagram (RBD), in which components are combined into blocks in series, in parallel or in the *k-out-of-n* configuration (which will be introduced in the following sections). We use the following example to illustrate the use of RBD.

Example 1.12

Consider the system shown in Figure 1.11, consisting of five stages, with $n_1 = n_2 = n_5 = 1$, and $n_3 = 3$ and $n_4 = 2$. Also

$$R_1 = 0.95, R_2 = 0.99, R_3 = 0.70, R_4 = 0.75, \text{ and } R_5 = 0.9.$$

Then

$$\begin{aligned} R_{sp} &= 0.95 \cdot 0.99 \cdot \left(1 - (1 - 0.7)^3\right) \cdot \left(1 - (1 - 0.75)^2\right) \cdot 0.9 \\ &= 0.772. \end{aligned}$$

♯

Fault trees provide another way to model system reliability [HENL 1981, MISR 1992, SAHN 1996]. A fault tree is a graphical representation of the combination of events that can cause the occurrence of system failure. An event is either a basic (primary) event or a logical combination of lower-level events. We assume that basic events are mutually independent and that probabilities for their occurrences are known. The occurrence of each event is denoted by a logic 1 at that node; otherwise the logic value of the node is 0. Logic value 1 for a basic event denotes failure of the corresponding component. Each gate has several inputs and one output. The inputs to a gate are either basic events or the outputs of other gates. The output of an **and** gate is a logic 1, if and only if, all of its inputs are logic 1. The output of an **or** gate is a logic 1 if one or more of its inputs are logic 1. There is a single output of the fault tree as a whole, called the *top event*, representing system failure.

Example 1.13

Consider a reliability model of alternate routing in a telephone network [BALA 1996]. The network is represented by a graph whose nodes denote the office locations

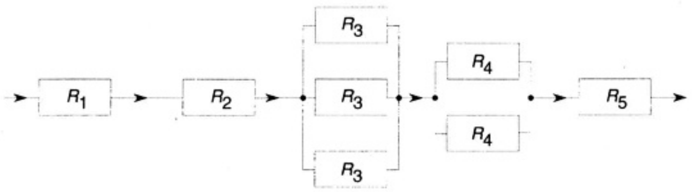


Figure 1.11. A series–parallel reliability block diagram

of a corporation and edges of the graph represent communication links between office locations as shown in Figure 1.12. The measure of interest is reliability, R, a measure of the network's ability to maintain a given set of connections. In Figure 1.12, the network is up whenever node-pairs A–B and C–D are both connected, either directly, or by the two-link alternate routes listed. We impose the condition that the alternate routes of the node pair A–B should be disjoint from those of node pair C–D. We assume that link failures are mutually independent. The fault tree is shown in Figure 1.13.

In a fault tree such as that in Figure 1.13, reliabilities of inputs to an or gate multiply while unreliabilities of inputs to an and gate multiply. Hence the network reliability is given by

$$\begin{aligned}R_{network} =& \left[1-(1-R_{ab})(1-R_{ac}R_{cb})(1-R_{ad}R_{db})(1-R_{ae}R_{eb})\right]\\ &\cdot\left[1-(1-R_{cd})(1-R_{ce}R_{ed})\right].\end{aligned}$$

♯

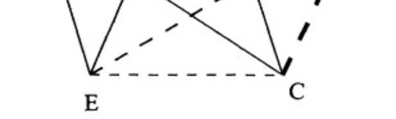
Reliability of systems with more general interconnections cannot be computed with the preceding formula. In such a case, we may obtain structure function [MISR 1992] of the system first, then compute the reliability of the system. The structure function of a system is defined as follows.

Definition (Structure Function). Let $\mathbf{X}$ be a state vector of a system with n components so that $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ where

$$x_i = \begin{cases} 1 & \text{if component } i \text{ is functioning,}\\ 0 & \text{if component } i \text{ has failed.}\end{cases}$$

The structure function $\Phi(\mathbf{X})$ is defined by

$$\Phi(\mathbf{X}) = \begin{cases} 1 & \text{if system is functioning,}\\ 0 & \text{if system has failed.}\end{cases}$$



Alternate routes

For C-D: C-E-D

For A-B: A-C-B, A-D-B, A-E-B

Figure 1.12. A communication network with five nodes

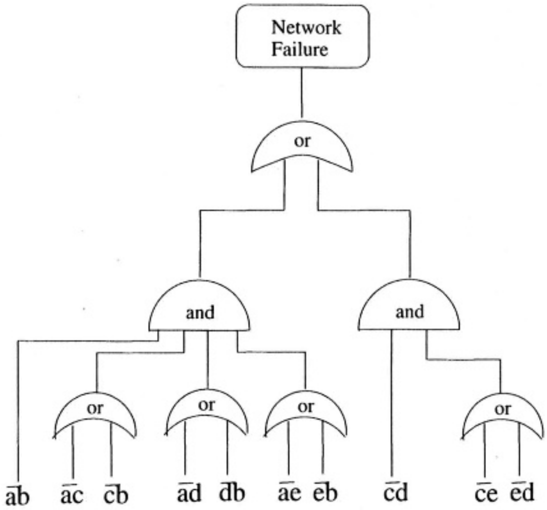


Figure 1.13. Fault tree for the communication network

Using the definition of system structure function, the reliability of a system can be written as

$$R = P\big(\Phi(\mathbf{X}) = 1\big).$$

Example 1.14

Consider the fault tree shown in Figure 1.14. Notice that event $\overline{B}_3$ is input to two gates; thus, the fault tree is said to have repeated (or shared) events. Such fault trees can no longer be solved by the simple method used for the fault tree without repeated events that we encountered in Example 1.13. For the current example, we have

$$\begin{aligned}\{\Phi = 0\} &= \big(\overline{A}_1 \cup (\overline{B}_1 \cap \overline{B}_3)\big) \cap \big(\overline{A}_2 \cup (\overline{B}_2 \cap \overline{B}_3)\big)\\ &= (\overline{A}_1 \cap \overline{A}_2) \cup (\overline{A}_1 \cap \overline{B}_2 \cap \overline{B}_3) \cup (\overline{A}_2 \cap \overline{B}_1 \cap \overline{B}_3) \cup (\overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3).\end{aligned}$$

Note that these four events are not mutually exclusive. Therefore, we cannot directly use axiom (A3), however, we could use SDP formula, i.e., relation (Re), to make them disjoint. Then, the reliability of the system is

$$\begin{aligned}R &= 1 - P(\Phi = 0)\\ &= 1 - P\big((\overline{A}_1 \cap \overline{A}_2) \cup (\overline{A}_1 \cap \overline{B}_2 \cap \overline{B}_3) \cup (\overline{A}_2 \cap \overline{B}_1 \cap \overline{B}_3) \cup (\overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3)\big)\\ &= 1 - P\big((\overline{A}_1 \cap \overline{A}_2) \cup (\overline{A}_1 \cap A_2 \cap \overline{B}_2 \cap \overline{B}_3) \cup (A_1 \cap \overline{A}_2 \cap \overline{B}_1 \cap \overline{B}_3)\\ &\quad \cup (A_1 \cap A_2 \cap \overline{B}_1 \cap \overline{B}_2 \cap \overline{B}_3)\big)\\ &= 1 - F_{A_1}F_{A_2} - F_{A_1}R_{A_2}F_{B_2}F_{B_3} - R_{A_1}F_{A_2}F_{B_1}F_{B_3} - R_{A_1}R_{A_2}F_{B_1}F_{B_2}F_{B_3}\end{aligned}$$

where $F_x = 1 - R_x$.

♯

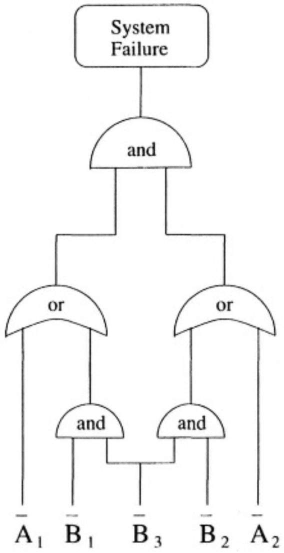


Figure 1.14. A fault tree

Starting with system structure function, there are two methods to obtain system reliability: (1) the use of inclusion–exclusion formula (Rd) and (2) the use of the SDP formula illustrated above. For an efficient implementation of the SDP method, see Luo and Trivedi [LUO 1998]. A third, even more efficient approach is based on the binary decision diagrams (BDDs) [ZANG 1999]. A fourth method is based on the use of conditioning (also called factoring) to be discussed in the next section. The BDD approach and factoring approach do not need the structure function to begin with. Further note that reliability of systems with standby redundancy cannot be computed using methods discussed in this chapter, but techniques to be discussed later in this book will enable us to do so.

Problems

1. Two towns are connected by a network of communication channels. The probability of a channel's failure-free operation is R, and channel failures are independent. Minimal level of communication between towns can be guaranteed provided at least one path containing properly functioning channels exists. Given the network of Figure 1.P.1, determine the probability that the two towns will be able to communicate. Here ⊣⊢ denotes a communication channel.

2. Given three components with respective reliabilities $R_1 = 0.8$, $R_2 = 0.75$, and $R_3 = 0.98$, compute the reliabilities of the three systems shown in Figure 1.P.2.

3. Determine the conditions under which an event A is independent of its subset B.

4. *General multiplication rule* (GMR). Given a list of events $A_1, A_2, \ldots, A_n$ (not

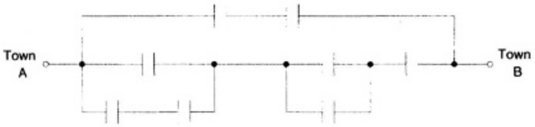


Figure 1.P.1. A network of communication channels

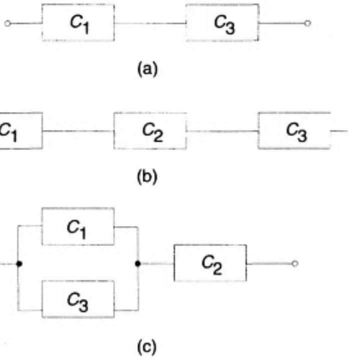


Figure 1.P.2. Reliability block diagrams

necessarily independent), show that

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_n) &= P[A_1|(A_2 \cap A_3 \cap \cdots \cap A_n)] \\ &\quad \cdot P[A_2|(A_3 \cap \cdots \cap A_n)] \\ &\quad \cdot P[A_3|(A_4 \cap \cdots \cap A_n)] \\ &\quad \cdots \\ &\quad \cdot P(A_{n-1}|A_n)P(A_n), \end{aligned}$$

provided all the conditional probabilities on the right-hand side are defined.

5. Seven lamps are located as shown in Figure 1.P.3. Each lamp can fail with probability q, independently of all the others. The system is operational if no two adjacent lamps fail. Obtain an expression for system reliability.

6. Consider a base repeater in a cellular communication system with two control channels and three voice channels. Assume that the system is up so long as at least one control channel and at least one voice channel is functioning. Draw a reliability block diagram for this problem and write down an expression for system reliability. Next, draw a fault tree model for this system. Note that this fault tree has no repeated events and hence can be solved in a way similar to that for a series–parallel reliability block diagram.

7. Modify the base repeater problem above so that a control can also function as a voice channel. Draw a fault tree model for the modified problem. Notice that the fault tree has repeated events. Derive the reliability expression using the SDP method.

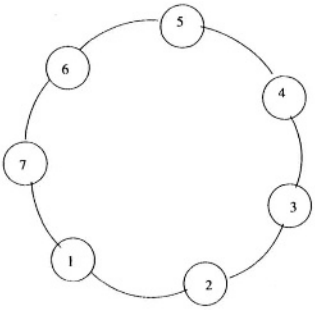
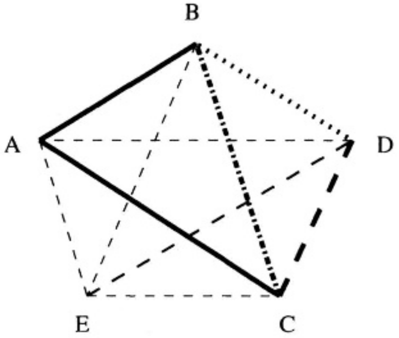


Figure 1.P.3. Lamp problem



Alternate routes

For C-D: C-B-D

For A-B: A-C-B

B-C is shared.

Figure 1.P.4. A modified communication network

8. Return to Example 1.13 but now permitting a shared link B–C as shown in Figure 1.P.4. Draw the fault tree for modeling the reliability for the communication network. Note that due to the shared link, the fault tree will have a shared or repeated event. Derive an expression for system reliability using SDP method as in Example 1.14.

1.11 BAYES' RULE

A given event B of probability $P(B)$ partitions the sample space S into two disjoint subsets B and $\overline{B}$. If we consider $S' = \{B, \overline{B}\}$ and associate the probabilities $P(B)$ and $P(\overline{B})$ to the respective points in S', then S' is very similar to a sample space, except that there is a many-to-one correspondence between the outcomes of the experiment and the elements of S'. A space such as S' is often called an **event space**. In general, a list of n events

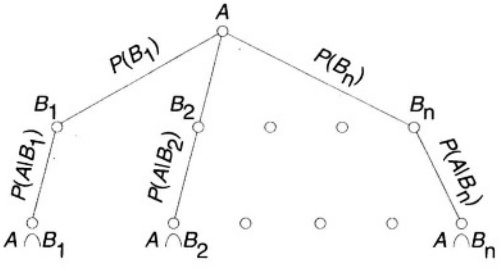


Figure 1.15. The theorem of total probability

$B_1, B_2, \ldots, B_n$ that are collectively exhaustive and mutually exclusive form an **event space**, $S' = \{B_1, B_2, \ldots, B_n\}$.

Returning to the event space $S' = \{B, \overline{B}\}$, note that an event A is partitioned into two disjoint subsets:

$$A = (A \cap B) \cup (A \cap \overline{B}).$$

Then by axiom (A3):

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \\ &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \end{aligned}$$

by definition of conditional probability.

This relation is analogous to Shannon's theorem in switching theory and can be generalized with respect to the event space $S' = \{B_1, B_2, \ldots, B_n\}$:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{1.11}$$

This relation is also known as the **theorem of total probability**, and is sometimes called the **rule of elimination**. This situation can be visualized by constructing a **tree diagram** (or a **probability tree**) as shown in Figure 1.15, where each branch is so labeled that the product of all branch probabilities from the root of the tree to any node equals the probability of the event represented by that node. Now $P(A)$ can be computed by summing probabilities associated with all the leaf nodes of the tree. In practice, after the experiment, a situation often arises in which the event A is known to have occurred, but it is not known directly which of the mutually exclusive and collectively exhaustive events $B_1, B_2, \ldots, B_n$ has occurred. In this situation, we may be interested in evaluating $P(B_j|A)$, the conditional probability that one of these events B_j occurs, given that A occurs. By applying the definition of conditional probability followed by the use of theorem of total probability, we find that

$$
\begin{aligned}
P(B_j|A) &= \frac{P(B_j \cap A)}{P(A)} \\
&= \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}.
\end{aligned}
\tag{1.12}
$$

This relation is known as *Bayes' rule* and is useful in many applications. This rule also forms the basis of a statistical method called **Bayesian procedure**. $P(B_j|A)$ is sometimes called an **a posteriori probability**.

Example 1.15

Measurements at the North Carolina Super Computing Center (NCSC) on a certain day, indicated that the source of incoming jobs is 15% from Duke, 35% from University of North Carolina (UNC), and 50% from North Carolina State (NC State). Suppose that the probabilities that a job initiated from these universities is a multitasking job are 0.01, 0.05, and 0.02, respectively. Find the probability that a job chosen at random at NCSC is a multitasking job. Also find the probability that a randomly chosen job comes from the University of North Carolina, given that it is a multitasking job.

Define the events B_i = "job is from university i" ($i = 1, 2, 3$ for Duke, UNC, and NC State, respectively), and A = "job uses multitasking." Then, by the theorem of total probability, we obtain

$$
\begin{aligned}
P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\
&= (0.01) \cdot (0.15) + (0.05) \cdot (0.35) + (0.02) \cdot (0.5) \\
&= 0.029.
\end{aligned}
$$

Now the second event of interest is $[B_2|A]$, and from Bayes' rule:

$$
\begin{aligned}
P(B_2|A) &= \frac{P(A|B_2)P(B_2)}{P(A)} \\
&= \frac{0.05 \cdot 0.35}{0.029} \\
&= 0.603.
\end{aligned}
$$

Note that the knowledge that the job uses multitasking increases the chance that it came from UNC from 35% to about 60%.

♯

Example 1.16

A binary communication channel carries data as one of two types of signals denoted by 0 and 1. As a result of noise, a transmitted 0 is sometimes received as a 1 and a transmitted 1 is sometimes received as a 0. For a given channel, assume a probability of 0.94 that a transmitted zero is correctly received as a zero and a probability of 0.91 that a transmitted one is received as a one. Further assume a probability of 0.45 of transmitting a 0. If a signal is sent, determine the

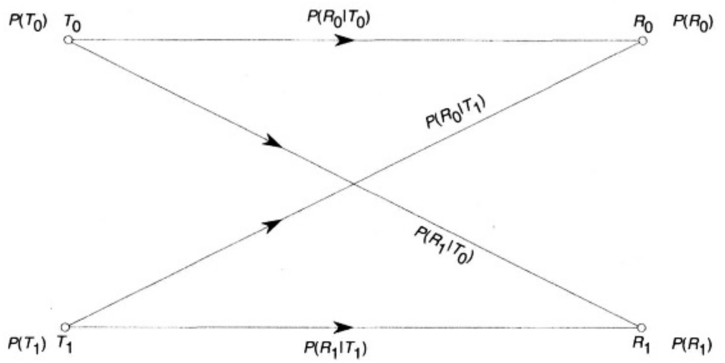


Figure 1.16. A channel diagram

1. Probability that a 1 is received.
2. Probability that a 0 is received.
3. Probability that a 1 was transmitted given that a 1 was received.
4. Probability that a 0 was transmitted given that a 0 was received.
5. Probability of an error.

Define events $T_0 =$ "a 0 is transmitted" and event $R_0 =$ "a 0 is received." Then let $T_1 = \overline{T}_0 =$ "a 1 is transmitted" and $R_1 = \overline{R}_0 =$ "a 1 is received." Then the events of interest under items 1, 2, 3, and 4 are respectively given by R_1, R_0, $[T_1|R_1]$, and $[T_0|R_0]$. An error in the transmitted signal is the union of the two disjoint events $[T_1 \cap R_0]$ and $[T_0 \cap R_1]$. The operation of a binary communication channel may be visualized by a **channel diagram** shown in Figure 1.16. In the given problem, we have $P(R_0|T_0) = 0.94$, $P(R_1|T_1) = 0.91$, and $P(T_0) = 0.45$. From these we get

$$\begin{aligned} P(R_1|T_0) &= P(\overline{R}_0|T_0) = 1 - P(R_0|T_0) = 0.06, \\ P(R_0|T_1) &= P(\overline{R}_1|T_1) = 1 - P(R_1|T_1) = 0.09, \\ P(T_1) &= P(\overline{T}_0) = 1 - P(T_0) = 0.55. \end{aligned}$$

Now from the theorem of total probability:

$$\begin{aligned} P(R_0) =& P(R_0|T_0)P(T_0) + P(R_0|T_1)P(T_1) \\ =& (0.94) \cdot (0.45) + (0.09) \cdot (0.55) \\ =& 0.423 + 0.0495 \\ =& 0.4725, \\ P(R_1) =& P(\overline{R}_0) \\ =& 1 - P(R_0) \\ =& 1 - 0.4725 \\ =& 0.5275. \end{aligned}$$

Using Bayes' rule, we have

$$\begin{aligned}P(T_1|R_1) =&\frac{P(R_1|T_1)P(T_1)}{P(R_1)}\\ =&\frac{0.91\cdot 0.55}{0.5275}\\ =&0.9488,\\ P(T_0|R_0) =&\frac{P(R_0|T_0)P(T_0)}{P(R_0)}\\ =&\frac{0.94\cdot 0.45}{0.4725}\\ =&0.8952.\end{aligned}$$

Now:

$$\begin{aligned}P(T_1|R_0) =&P(\overline{T}_0|R_0)\\ =&1-P(T_0|R_0)\\ =&0.1048,\\ P(T_0|R_1) =&1-P(T_1|R_1)\\ =&0.0512\end{aligned}$$

and

$$\begin{aligned}P(\text{“error”}) =&P(T_1\cap R_0)+P(T_0\cap R_1)\\ =&P(T_1|R_0)P(R_0)+P(T_0|R_1)P(R_1)\\ =&0.1048\cdot 0.4725+0.0512\cdot 0.5275\\ =&0.0765.\end{aligned}$$

Alternately, the error probability can be evaluated by

$$\begin{aligned}P(\text{“error”}) =&P(T_1\cap R_0)+P(T_0\cap R_1)\\ =&P(R_0|T_1)P(T_1)+P(R_1|T_0)P(T_0)\\ =&0.09\cdot 0.55+0.06\cdot 0.45=0.0765.\end{aligned}$$

[*Quiz*: Construct an appropriate sample space for this problem.]

♯

Example 1.17

A given lot of VLSI chips contains 2% defective chips. Each chip is tested before delivery. The tester itself is not totally reliable so that

$$\begin{aligned}P(\text{“tester says chip is good”}|\text{“chip is actually good”}) &= 0.95,\\ P(\text{“tester says chip is defective”}|\text{“chip is actually defective”}) &= 0.94.\end{aligned}$$

If a tested device is indicated to be defective, what is the probability that it is actually defective?

By Bayes' rule, we have

$$\begin{aligned}
&P(\text{“chip is defective”}|\text{“tester says it is defective”})\\
&= \frac{P(\text{“tester says defective”}|\text{“chip defective”})P(\text{“chip defective”})}{\begin{array}{l}P(\text{“tester says defective”}|\text{“chip defective”})P(\text{“chip defective”})\\ +P(\text{“tester says defective”}|\text{“chip is good”})P(\text{“chip is good”})\end{array}}\\
&= \frac{0.94 \cdot 0.02}{0.94 \cdot 0.02 + 0.05 \cdot 0.98}\\
&= \frac{0.0188}{0.0188 + 0.049}\\
&= \frac{0.0188}{0.0678}\\
&= 0.2772861.
\end{aligned}$$

♯

Example 1.18

We have seen earlier how to compute the reliability of series–parallel systems. However, many systems in practice do not conform to a series–parallel structure. As an example, consider evaluating the reliability R of the five-component system shown in Figure 1.17. The system is said to be functioning properly only if all the components on at least one path from point A to point B are functioning properly.

Define for $i = 1, 2, \ldots, 5$ event X_i = "component i is functioning properly," and let R_i = reliability of component $i = P(X_i)$. Let X = "system functioning properly" and let R = "system reliability" $= P(X)$. It is clear that X is union of four events:

$$X = (X_1 \cap X_4) \cup (X_2 \cap X_4) \cup (X_2 \cap X_5) \cup (X_3 \cap X_5). \tag{1.13}$$

These four events are not mutually exclusive. Therefore, we cannot directly use axiom (A3). Note, however, that we could use relation (Rd), which does apply to a union of intersecting events. But this method is computationally tedious for a relatively long list of events. We could use the sum of disjoint products (SDP)

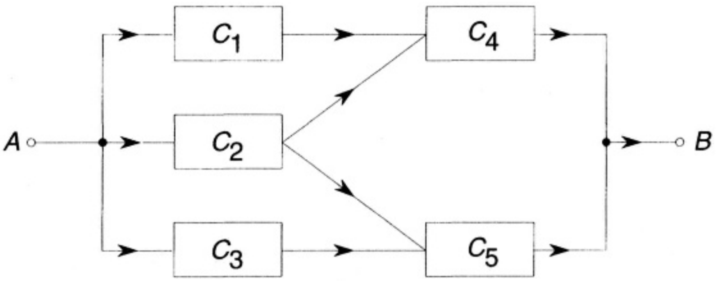


Figure 1.17. A non-series–parallel system

method (Relation Re) in this case. We illustrate the use of yet another method known as factoring or conditioning in this case. Observe that using the theorem of total probability, we have

$$\begin{aligned} P(X) &= P(X|X_2)P(X_2) + P(X|\overline{X}_2)P(\overline{X}_2) \\ &= P(X|X_2)R_2 + P(X|\overline{X}_2)(1 - R_2). \end{aligned} \tag{1.14}$$

Now to compute $P(X|X_2)$, we observe that since component C_2 is functioning, the status of components C_1 and C_3 are irrelevant. Thus, under this condition, the system is equivalent to two components C_4 and C_5 in parallel. Therefore using formula (1.9) we get

$$P(X|X_2) = 1 - (1 - R_4)(1 - R_5). \tag{1.15}$$

To compute $P(X|\overline{X}_2)$, we observe that since component C_2 is known to have malfunctioned, the resulting equivalent system is a series–parallel one whose reliability is easily computed:

$$P(X|\overline{X}_2) = 1 - (1 - R_1R_4)(1 - R_3R_5). \tag{1.16}$$

Combining equations (1.14)–(1.16), we have

$$\begin{aligned} R &= \big[1 - (1 - R_4)(1 - R_5)\big]R_2 + \big[1 - (1 - R_1R_4)(1 - R_3R_5)\big](1 - R_2) \\ &= 1 - R_2(1 - R_4)(1 - R_5) - (1 - R_2)(1 - R_1R_4)(1 - R_3R_5). \end{aligned}$$

♯

Problems

1. A technique for fault-tolerant software, suggested by Randell [RAND 1978], consists of a primary and an alternate module for each critical task, together with a test for determining whether a module performed its function correctly. Such a construct is called a **recovery block**. Define the following events:

   A = "primary module functions correctly."
   B = "alternate module functions correctly."
   D = "detection test following the execution of the primary
   performs its task correctly."

   Assume that event pairs A and D as well as B and D are independent but events A and B are dependent. Derive an expression for the failure probability of a recovery block [HECH 1976]. (*Hint*: Use a tree diagram.)

2. Consider the non-series–parallel system of four independent components shown in Figure 1.P.5. The system is considered to be functioning properly if all components along at least one path from input to output are functioning properly. Determine an expression for system reliability as a function of component reliabilities. Also draw an equivalent fault tree model for the reliability block diagram described above.

3. A lot of components contains 0.6% defectives. Each component is subjected to a test that correctly identifies a defective, but about 2 in every 100 good components is also indicated defective. Given that a randomly chosen component is declared defective by the tester, compute the probability that it is actually defective.

4. A certain firm has plants A, B, and C producing respectively 35%, 15%, and 50%, of the total output. The probabilities of a nondefective product are, respectively, 0.75, 0.95, and 0.85. A customer receives a defective product. What is the probability that it came from plant C?

5. Consider a trinary communication channel [STAR 1979] whose channel diagram is shown in Figure 1.P.6. For $i = 1, 2, 3$ let T_i denote the event "digit i is transmitted" and let R_i denote the event "digit i is received." Assume that a 3 is transmitted 3 times more frequently than a 1, and a 2 is sent twice as often as 1. If a 1 has been received, what is the expression for the probability that a 1 was sent? Derive an expression for the probability of a transmission error.

6. Of all the graduate students in a university, 70% are women and 30% are men. Suppose that 20% and 25% of the female and male population, respectively, smoke cigarettes. What is the probability that a randomly selected graduate student is

   (a) A woman who smokes?

   (b) A man who smokes?

   (c) A smoker?

7. Compute the reliability of the system discussed in Example 1.18 (Figure 1.17), starting from equation (1.13), first using the inclusion-exclusion formula (Rd) and then using the SDP formula (Re). Also draw the fault tree model of this system.

8. Yet another method of evaluating the reliability of the system such as that discussed in Example 1.16 is to use the methods of switching theory. Noting that X_1, X_2, X_3, X_4, X_5 are Boolean variables and X is a switching function of these variables, we can draw a truth table with $2^5 = 32$ rows. Rows of the truth table represent a collection of mutually independent and collectively exhaustive events. Each row represents an elementary event that is an intersection of independent

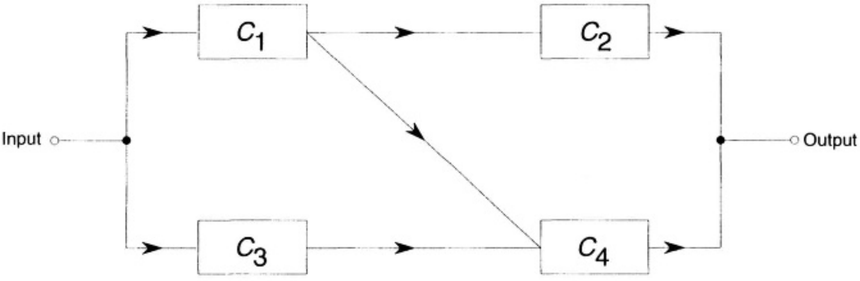


Figure 1.P.5. Another non-series–parallel system

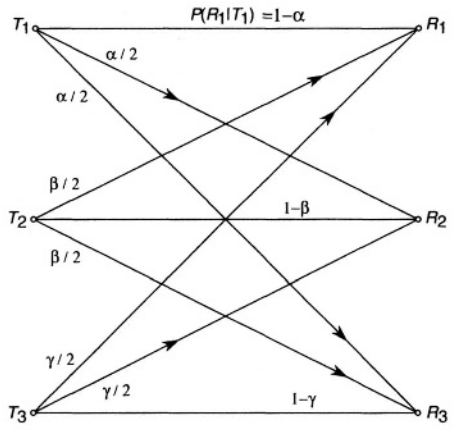


Figure 1.P.6. A trinary communication channel: channel diagram

events and hence its probability can be computed. For example, the elementary event $\overline{X}_1 \cap X_2 \cap \overline{X}_3 \cap X_4 \cap X_5$ is assigned the probability $(1-R_1)R_2(1-R_3)R_4R_5$. Computing $P(X)$ now reduces to adding up probabilities of rows of the truth table with 1s in the function column. Use this method to compute the reliability of the system in Figure 1.17. This method is called the **state enumeration method** or the **Boolean truth table method**.

1.12 BERNOULLI TRIALS

Consider a random experiment that has two possible outcomes, "success" and "failure" (or "hit" and "miss," or "good" and "defective," or "digit received correctly" and "digit received incorrectly") or the like. Let the probabilities of the two outcomes be p and q, respectively, with $p+q=1$. Now consider the compound experiment consisting of a sequence of n independent repetitions of this experiment. Such a sequence is known as a **sequence of Bernoulli trials**. This abstract sequence models many physical situations of interest to us:

1. Observe n consecutive executions of an **if** statement, with success = "**then** clause is executed" and failure = "**else** clause is executed."
2. Examine components produced on an assembly line, with success = "acceptable" and failure = "defective."
3. Transmit binary digits through a communication channel, with success = "digit received correctly" and failure = "digit received incorrectly."
4. Consider a computer system that allocates a finite quantum (or time slice) to a job scheduled for processor service, in an attempt to give fast service to requests for trivial processing. Observe n time slice terminations, with success = "job has completed processing" and failure =

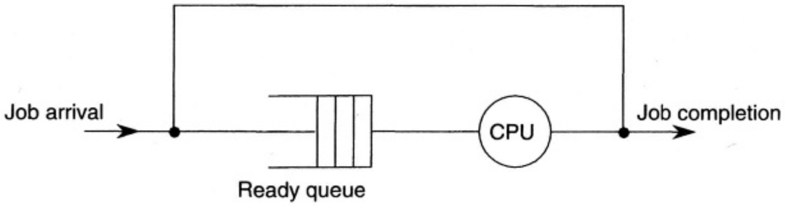


Figure 1.18. A CPU queue with time slicing

> "job still requires processing and joins the tail end of the ready queue of processes." This situation may be depicted as in Figure 1.18.

Let 0 denote failure and 1 denote success. Let S_n be the sample space of an experiment involving n Bernoulli trials, defined by

$$\begin{aligned} S_1 &= \{0,1\}, \\ S_2 &= \{(0,0),(0,1),(1,0),(1,1)\}, \\ S_n &= \{2^n \text{ } n\text{-tuples of 0s and 1s}\}. \end{aligned}$$

The probability assignment over the sample space S_1 is already specified: $P(0) = q \geq 0, P(1) = p \geq 0$, and $p + q = 1$. We wish to assign probabilities to the points in S_n.

Let A_i = "success on trial i" and $\overline{A}_i$ = "failure on trial i," then $P(A_i) = p$ and $P(\overline{A}_i) = q$. Now consider s an element of S_n such that $s = (1,1,\ldots,1,0,0,\ldots,0)$ (k 1s and $(n-k)$ 0s). Then the elementary event $\{s\}$ can be written as

$$\{s\} = A_1 \cap A_2 \cdots \cap A_k \cap \overline{A}_{k+1} \cap \cdots \cap \overline{A}_n$$

and

$$\begin{aligned} P(s) &= P(A_1 \cap A_2 \cdots \cap A_k \cap \overline{A}_{k+1} \cap \cdots \cap \overline{A}_n) \\ &= P(A_1)P(A_2) \cdots P(A_k)P(\overline{A}_{k+1}) \cdots P(\overline{A}_n) \end{aligned}$$

by independence. Therefore

$$P(s) = p^k q^{n-k}. \qquad (1.17)$$

Similarly, any sample point with k 1s and $(n-k)$ 0s is assigned probability $p^k q^{n-k}$. Noting that there are $\binom{n}{k}$ such sample points, the probability of obtaining exactly k successes in n trials is

$$p(k) = \binom{n}{k} p^k q^{n-k}, k = 0, 1, \ldots, n. \qquad (1.18)$$

We may verify that (1.18) is a legitimate probability assignment over the sample space S_n since

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} &= (p+q)^n\\ &= 1\end{aligned}$$

by the binomial theorem.

Consider the set of events $\{B_0, B_1, \ldots, B_n\}$ where $B_k = \{s \in S_n$ such that s has exactly k 1s and $(n-k)$ 0s$\}$. It is clear that this is a mutually exclusive and collectively exhaustive family of events. Furthermore

$$P(B_k) = \binom{n}{k}p^k q^{n-k} \geq 0 \text{ and } \sum_{k=0}^{n} P(B_k) = 1.$$

Therefore, this collection of events is an event space with $(n+1)$ events. Compare this with 2^n sample points in S_n. Thus, when in a physical situation, if we are concerned not with the actual sequence of successes and failures but merely with the number of successes and the number of failures, it is profitable to use the event space rather than the original sample space.

Example 1.19

Consider a binary communication channel transmitting coded words of n bits each. Assume that the probability of successful transmission of a single bit is p (and the probability of an error is $q = 1 - p$), and that the code is capable of correcting up to e (where $e \geq 0$) errors. For example, if no coding or parity checking is used, then $e = 0$. If a single error correcting Hamming code is used then $e = 1$. For more details on this topic, see Hamming [HAMM 1980]. If we assume that the transmission of successive bits is independent, then the probability of successful word transmission is

$$\begin{aligned}P_w &= P(\text{“}e\text{ or fewer errors in }n\text{ trials”})\\ &= \sum_{i=0}^{e}\binom{n}{i}(1-p)^i p^{n-i}.\end{aligned}$$

♯

Example 1.20

In connection with reliability computation, we have considered series and parallel systems. Now we consider a system with n components that requires k $(\leq n)$ or more components to function for the correct operation of the system. Such systems are often called k-out-of-n systems. If we let $k = n$, then we have a series system; if we let $k = 1$, then we have a system with parallel redundancy. Assume that all n components are statistically identical and function independently of each other. If we let R denote the reliability of a component (and $q = 1 - R$ gives its unreliability), then the experiment of observing the statuses of n components can be thought of as a sequence of n Bernoulli trials with the probability of success equal to R. Now the reliability of the system is

$$\begin{aligned}
R_{k|n} &= P(\text{“}k \text{ or more components functioning properly”}) \\
&= P(\bigcup_{i=k}^{n}\{\text{“exactly } i \text{ components functioning properly”}\}) \\
&= \sum_{i=k}^{n} P(\text{“exactly } i \text{ components functioning properly”}) \\
&= \sum_{i=k}^{n} p(i), \\
R_{k|n} &= \sum_{i=k}^{n} \binom{n}{i} R^i (1-R)^{n-i} \qquad (1.19)
\end{aligned}$$

Verify that $R_{1|n} = R_p$:

$$\begin{aligned}
R_{1|n} &= \sum_{i=1}^{n} \binom{n}{i} R^i (1-R)^{n-i} \\
&= \sum_{i=0}^{n} \binom{n}{i} R^i (1-R)^{n-i} - \binom{n}{0} R^0 (1-R)^n \\
&= [R + (1-R)]^n - (1-R)^n \\
&= 1 - (1-R)^n.
\end{aligned}$$

Verify that $R_{n|n} = R_s$:

$$\begin{aligned}
R_{n|n} &= \sum_{i=n}^{n} \binom{n}{i} R^i (1-R)^{n-i} \\
&= \binom{n}{n} R^n (1-R)^0 \\
&= R^n
\end{aligned}$$

♯

As another special case of formula (1.19), consider a system with triple modular redundancy, often known as TMR or a triplex system (see Figure 1.19). In such a system there are three components, two of which are required to be in working order for the system to function properly (i.e., $n = 3$ and $k = 2$). This is achieved by feeding the outputs of the three components into a majority voter. Then

$$\begin{aligned}
R_{\text{TMR}} &= \sum_{i=2}^{3} \binom{3}{i} R^i (1-R)^{(3-i)} \\
&= \binom{3}{2} R^2 (1-R) + \binom{3}{3} R^3 (1-R)^0 \\
&= 3R^2(1-R) + R^3
\end{aligned}$$

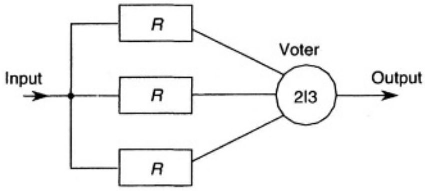


Figure 1.19. A triple modular redundant system

and thus

$$R_{\text{TMR}} = 3R^2 - 2R^3. \tag{1.20}$$

Note that

$$R_{\text{TMR}} = \begin{cases} > R & \text{if } R > \frac{1}{2}, \\ = R & \text{if } R = \frac{1}{2}, \\ < R & \text{if } R < \frac{1}{2}. \end{cases}$$

Thus TMR increases reliability over the simplex system only if the simplex reliability is greater than 0.5; otherwise this type of redundancy actually *decreases* reliability.

It should be noted that the voter output simply corresponds to the majority, and therefore it is possible for two or more malfunctioning units to agree, producing an erroneous voter output. Additional detection logic is required to avoid this situation. Also, the unreliability of the voter will further degrade the TMR reliability.

In the above example, we assumed that the n successive trials have the same probability of success. Now consider **nonhomogeneous Bernoulli trials**, where probability of success changes with each trial. In the reliability context, let R_i denote the reliability of the ith component for $i = 1, ..., n$. Then the calculation is a bit more complicated [SAHN 1996]:

$$R_{k|n} = 1 - \sum_{|I| \geq k} \left(\prod_{i \in I} (1 - R_i) \right) \left(\prod_{i \notin I} R_i \right), \tag{1.21}$$

where I ranges over all choices $i_1 < i_2 < ... < i_m$ such that $k \leq m \leq n$.

Let us still consider the TMR system with $n = 3$ and $k = 2$. However, the individual reliabilities are not identical any longer. Then, by formula (1.21), we have

$$\begin{aligned} R_{2|3} &= 1 - (1 - R_1)(1 - R_2)R_3 - R_1(1 - R_2)(1 - R_3) \\ &\quad -(1 - R_1)R_2(1 - R_3) - (1 - R_1)(1 - R_2)(1 - R_3) \\ &= R_1R_2 + R_1R_3 + R_2R_3 - 2R_1R_2R_3 \end{aligned} \tag{1.22}$$

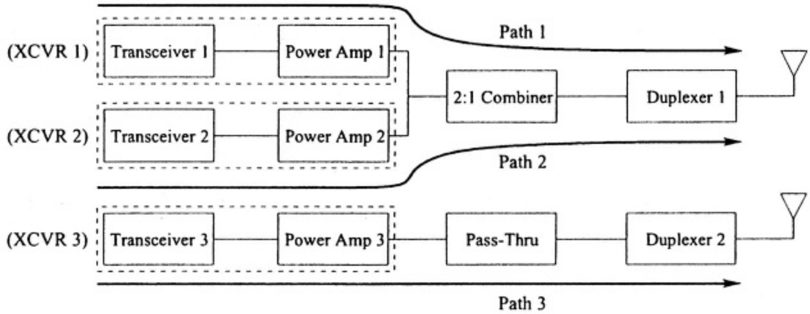


Figure 1.20. BTS sector/transmitter

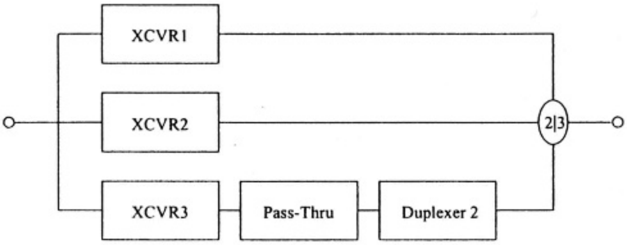


Figure 1.21. Reliability block diagram when *2: 1 combiner* and *duplexer 1* are up

Example 1.21 [DOSS 2000]

Consider a BTS (base transceiver system) sector/transmitter system shown in Figure 1.20. It consists of three RF (radio frequency) carriers (transceiver and power amplifier) on two antennas. In order for the system to be operational, at least two functional transmitter paths are needed.

We use the factoring method to arrive at the reliability block diagram for the system. Observe that the failure of the *2: 1 combiner* or *duplexer 1* would disable both path 1 and path 2, which would lead to system failure. So, we condition on these components. When both these components are functional, the system reliability is given by the RBD shown in Figure 1.21. As noted before, failure of any one of these two components results in system failure. Hence, the overall system reliability is captured by the RBD shown in Figure 1.22. If we let R_x, R_p, R_d, and R_c be the reliabilities of an XCVR, a pass-thru, a duplexer, and a combiner, then the reliabilities of XCVR1, XCVR2, XCVR3 with the "pass-thru" and duplexer 2, and the 2:1 combiner with duplexer 1 are $R_1 = R_x$, $R_2 = R_x$, $R_3 = R_x R_p R_d$, and $R_4 = R_c R_d$, respectively. Therefore, by formula (1.22), the overall system reliability is given by

$$\begin{aligned} R &= (R_1 R_2 + R_1 R_3 + R_2 R_3 - 2R_1 R_2 R_3) R_4 \\ &= (1 + 2R_p R_d - 2R_x R_p R_d) R_x^2 R_c R_d \end{aligned}$$

For a detailed discussion of various SDP methods and the factoring method of

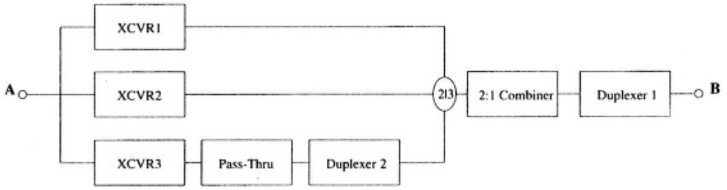


Figure 1.22. System reliability block diagram

reliability computation see Rai et al. [RAI 1995].

♯

Next, we consider **generalized Bernoulli trials**. Here we have a sequence of n independent trials, and on each trial the result is exactly one of the k possibilities $b_1, b_2, \ldots, b_k$. On a given trial, let b_i occur with probability $p_i, i = 1, 2, \ldots, k$ such that

$$p_i \geq 0 \text{ and } \sum_{i=1}^{k} p_i = 1.$$

The sample space S consists of all k^n n-tuples with components $b_1, b_2, \ldots, b_k$. To a point $s \in S$

$$s = (\underbrace{b_1, b_1, \ldots, b_1}_{n_1}, \underbrace{b_2, b_2, \ldots, b_2}_{n_2}, \ldots, \underbrace{b_k, \ldots, b_k}_{n_k})$$

we assign the probability of $p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where $\sum_{i=1}^{k} n_i = n$. This is the probability assigned to any n-tuple having n_i occurrences of b_i, where $i = 1, 2, \ldots, k$. The number of such n-tuples are given by the multinomial coefficient [LIU 1968]:

$$\begin{pmatrix} & n & \\ n_1 \; n_2 & \cdots & n_k \end{pmatrix} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

As before, the probability that b_1 will occur n_1 times, b_2 will occur n_2 times, $\ldots$, and b_k will occur n_k times is given by

$$p(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{1.23}$$

and

$$\begin{aligned} \sum_{n_i \geq 0} p(n_1, n_2, \ldots, n_k) &= (p_1 + p_2 + \cdots + p_k)^n \\ &= 1 \end{aligned}$$

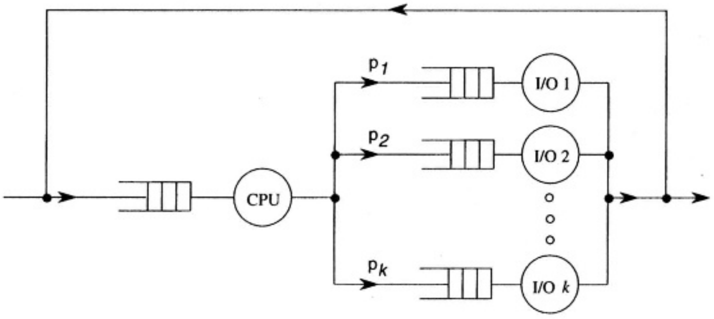


Figure 1.23. A CPU to I/O device queuing scheme

(where $\sum n_i = n$) by the multinomial theorem.

If we let $k = 2$, then generalized Bernoulli trials reduce to ordinary Bernoulli trials where $b_1 =$ "success," $b_2 =$ "failure," $p_1 = p$, $p_2 = q = 1 - p$, $n_1 = k$, and $n_2 = n - k$.

Two situations of importance are examples of generalized Bernoulli trials:

1. We are given that at the end of a CPU (central processing unit) burst, a program will request service from an I/O device i with probability p_i, where $i = 1, 2, \ldots, k$ and $\sum_i p_i = 1$. If we assume that successive CPU bursts are independent of each other, then the observation of n CPU burst terminations corresponds to a sequence of generalized Bernoulli trials. This situation may be pictorially visualized by the queuing network shown in Figure 1.23.

2. If we observe n consecutive independent executions of a switch statement (see below), then we have a sequence of generalized Bernoulli trials where p_i is the probability of executing the statement group S_i on an individual trial.

```
switch( I ) {

case 1: S1;

case 2: S2;

        ⋮

case k: Sk;

}
```

Example 1.22

Out of every 100 jobs received at a server, 50 are of class 1, 30 of class 2, and 20 of class 3. A sample of 30 jobs is taken with replacement.

1. Find the probability that the sample will contain 10 jobs of each class.
2. Find the probability that there will be exactly 12 jobs of class 2.

This is an example of generalized Bernoulli trials with $k = 3$, $n = 30$, $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. The answer to part (1) is

$$\begin{aligned} p(10,10,10) &= \frac{30!}{10! \cdot 10! \cdot 10!} \cdot 0.5^{10} \cdot 0.3^{10} \cdot 0.2^{10} \\ &= 0.003278. \end{aligned}$$

The answer to part (2) is obtained more easily if we collapse class 1 and class 3 together and consider this as an example of an ordinary Bernoulli trial with $p = 0.3$ (success corresponds to a class 2 job), $q = 1 - p = 0.7$ (failure corresponds to a class 1 or class 3 job). Then the required answer is as follows:

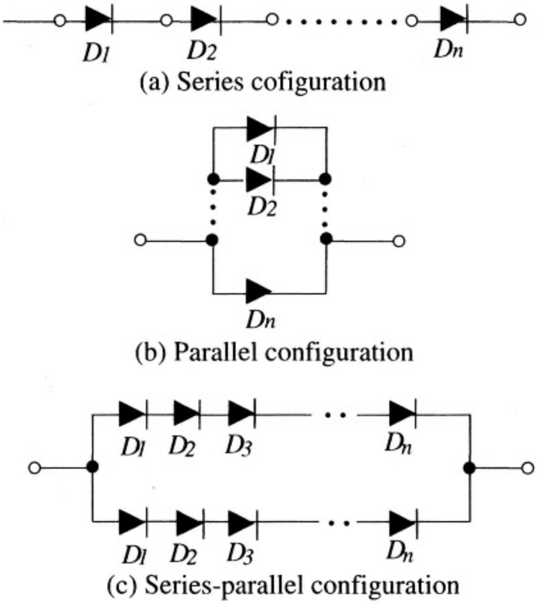
$$\begin{aligned} p(12) &= \binom{30}{12} \cdot 0.3^{12} \cdot 0.7^{18} \\ &= \frac{30!}{12! \cdot 18!} \cdot 0.3^{12} \cdot 0.7^{18} \\ &= 0.07485. \end{aligned}$$

♯

Example 1.23

So far, we have assumed that a component is either functioning properly or it has malfunctioned. Sometimes it is useful to consider more than two states. For example, a diode functions properly with probability p_1, develops a short circuit with probability p_2, and develops an open circuit with probability p_3 such that $p_1 + p_2 + p_3 = 1$. Thus there are two types of malfunctions, an open circuit and a closed circuit. In order to protect against such malfunctions, we investigate three types of redundancy schemes (refer to Figure 1.24): (a) a series connection, (b) a parallel connection, and (c) a series–parallel configuration.

First we analyze the series configuration. Let s_1, s_2, and s_3 respectively denote the probabilities of correct functioning, a short circuit, and an open circuit for the series configuration as a whole. The experiment of observing n diodes corresponds to a sequence of n generalized Bernoulli trials. Let n_1 diodes be functioning properly, n_2 diodes be short-circuited, and n_3 diodes be open-circuited. Then the event "the series configuration is functioning properly" is described by "none of the diodes is open-circuited and at least one of the diodes is functioning properly." This event consists of the sample points $\{(n_1, n_2, n_3) | n_1 \geq 1, n_2 \geq 0, n_3 = 0, n_1 + n_2 = n\}$.



(a) Series cofiguration

(b) Parallel configuration

(c) Series-parallel configuration

Figure 1.24. (a) Series configuration; (b) parallel configuration; (c) series–parallel configuration

Therefore

$$\begin{aligned} s_1 &= \sum_{\substack{n_1 \geq 1 \\ n_2 \geq 0 \\ n_1+n_2=n}} p(n_1, n_2, 0) \\ &= \sum \binom{n}{n_1, n_2, 0} p_1^{\ n_1} p_2^{\ n_2} p_3^{\ 0} \\ &= \sum_{n_1 \geq 1} \frac{n!}{n_1!(n-n_1)!} p_1^{\ n_1} p_2^{\ n-n_1} \\ &= \sum_{n_1=0}^{n} \binom{n}{n_1} p_1^{\ n_1} p_2^{\ n-n_1} - \frac{n!}{0!n!} p_1^{\ 0} p_2^{\ n} \\ &= (p_1 + p_2)^n - p_2^{\ n} \\ &= (1-p_3)^n - p_2^{\ n}. \end{aligned}$$

Note that $(1-p_3)^n$ is the probability that none of the diodes is open and $p_2^{\ n}$ is the probability that all diodes are short-circuited. Similarly

$$\begin{aligned} s_2 &= P(\text{“Series combination is short-circuited”}) \\ &= P(\text{“All diodes are short-circuited”}) \\ &= P(\{(n_1, n_2, n_3) | n_2 = n\}) \\ &= p_2^{\ n}. \end{aligned}$$

Also

$$\begin{aligned}
s_3 &= P(\text{“Series combination is open-circuited”}) \\
&= P(\text{“At least one diode is open-circuited”}) \\
&= p(\{(n_1, n_2, n_3) | n_3 \geq 1, n_1 + n_2 + n_3 = n\}) \\
&= 1 - P(\{(n_1, n_2, n_3) | n_3 = 0, n_1 + n_2 = n\}) \\
&= 1 - \sum_{n_1 + n_2 = n} \binom{n}{n_1, n_2} {p_1}^{n_1} {p_2}^{n_2} \\
&= 1 - (p_1 + p_2)^n \\
&= 1 - (1 - p_3)^n \\
&= 1 - P(\text{“no diodes are open-circuited”}).
\end{aligned}$$

Check that $s_1 + s_2 + s_3 = 1$.

Next, consider the parallel configuration, with P_i $(i = 1, 2, 3)$ respectively denoting the probabilities of properly functioning, short-circuit, and open-circuit situations. Then,

$$\begin{aligned}
P_1 &= P(\text{“parallel combination working properly”}) \\
&= P(\text{“at least one diode functioning and none of them short-circuited”}) \\
&= P(\{(n_1, n_2, n_3) | n_1 \geq 1, n_2 = 0, n_1 + n_3 = n\}) \\
&= (1 - p_2)^n - {p_3}^n \\
&= P(\text{“no diodes short-circuited”}) - P(\text{“all diodes are open-circuited”}), \\
P_2 &= P(\{(n_1, n_2, n_3) | n_2 \geq 1, n_1 + n_2 + n_3 = n\}) \\
&= 1 - (1 - p_2)^n, \\
P_3 &= P(\{(n_1, n_2, n_3) | n_3 = n\}) \\
&= {p_3}^n
\end{aligned}$$

To analyze the series–parallel configuration, we first reduce each one of the series configurations to an “equivalent” diode with respective probabilities s_1, s_2, and s_3. The total configuration is then a parallel combination of two “equivalent” diodes. Thus the probability that series–parallel diode configuration functions properly is given by

$$\begin{aligned}
R_1 &= (1 - s_2)^2 - {s_3}^2 \\
&= {s_1}^2 + 2 s_1 s_3 \\
&= s_1 (s_1 + 2 s_3) \\
&= [(1 - p_3)^n - {p_2}^n][(1 - p_3)^n - {p_2}^n + 2 - 2(1 - p_3)^n] \\
&= [(1 - p_3)^n - {p_2}^n][2 - (1 - p_3)^n - {p_2}^n].
\end{aligned}$$

♯

For an example of use of this technique in the context of availability analysis of VAXcluster systems, see Ibe et al. [IBE 1989]. For further study of multistate components (as opposed to two-state or binary components) and their reliability analysis, see Zang et al. [ZANG 1999].

Problems

1. Consider the following program segment:

```
if $B$ then
    repeat $S_1$ until $B_1$
else
    repeat $S_2$ until $B_2$
```

Assume that $P(B = \text{true}) = p, P(B_1 = \text{true}) = \frac{3}{5}$, and $P(B_2 = \text{true}) = \frac{2}{5}$. Exactly one statement is common to statement groups S_1 and S_2: write ("good day"). After many repeated executions of the preceding program segment, it has been estimated that the probability of printing exactly three "good day" messages is $\frac{3}{25}$. Derive the value of p.

2. Given that the probability of error in transmitting a bit over a communication channel is 8×10^{-4}, compute the probability of error in transmitting a block of 1024 bits. Note that this model assumes that bit errors occur at random, but in practice errors tend to occur in bursts. Actual block error rate will be considerably lower than that estimated here.

3. In order to increase the probability of correct transmission of a message over a noisy channel, a *repetition* code is often used. Assume that the "message" consists of a single bit, and that the probability of a correct transmission on a single trial is p. With a repetition code of rate $1/n$, the message is transmitted a fixed number (n) of times and a majority voter at the receiving end is used for decoding. Assuming $n = 2k + 1$, $k = 0, 1, 2 \ldots$, determine the error probability P_e of a repetition code as a function of k.

4. An application requires that at least two processors in a multiprocessor system be available with more than 95% probability. The cost of a processor with 60% reliability is \$1000, and each 10% increase in reliability will cost \$800. Determine the number of processors (n) and the reliability (p) of each processor (assume that all processors have the same reliability) that minimizes the total system cost.

5. Show that the number of terms in the multinomial expansion:

$$\left[\sum_{i=1}^{k}(p_i)\right]^n \text{ is } \binom{n+k-1}{n}.$$

Note that the required answer is the number of unordered sets of size n chosen from a set of k distinct objects with repetition allowed [LIU 1968].

6. A communication channel receives independent pulses at the rate of 12 pulses per microsecond (12 μs^{-1}). The probability of a transmission error is 0.001 for each pulse. Compute the probabilities of

   (a) No errors per microsecond
   (b) One error per microsecond
   (c) At least one error per microsecond
   (d) Exactly two errors per microsecond

7. Plot the reliabilities of a k out of n system as a function of the simplex reliability R $(0 \leq R \leq 1)$ using $n = 3$ and $k = 1, 2, 3$ [parallel redundancy, TMR (triple modular redundancy), and a series system, respectively].

8. Determine the conditions under which diode configurations in Figures 1.24(a)–(c) will improve reliability over that of a single diode. Use $n = 2$ to simplify the problem.

9. Consider a system with n capacitors in parallel. For the system to function properly, at least k-out-of-n capacitors should be functioning properly. A capacitor can fail in two modes: open and short (circuit). If a capacitor develops an open circuit, and the number of remaining working capacitors is greater than or equal to k, then the system still functions properly. If any one capacitor develops a short circuit then the system fails immediately. Given the probability of a capacitor functioning properly p_1=0.3, the probability of a capacitor developing a short circuit p_2=0.4, the probability of a capacitor developing an open circuit p_3=0.3, n=10 and k=7, calculate the probability of the system functioning properly.

10. Consider an example of n nonhomogeneous Bernoulli trials where a failure can occur on each trial independently, with a probability $1 - e^{-\alpha i}$ for the ith trial [KOVA 2000]. Prove that over n trials,

    (a) $P(\text{“no failure occurs”}) = e^{-[n(n+1)/2]\alpha}$

    (b) $P(\text{“no more than one failure occurs”}) = e^{-[n(n+1)/2]\alpha}\left[\frac{e^{\alpha}-e^{(n+1)\alpha}}{1-e^{\alpha}} - n + 1\right]$.

Review Problems

1. In the computation of TMR reliability, we assumed that when two units have failed they will both produce incorrect results and, hence after voting, the wrong answer will be produced by the TMR configuration. In the case that the two faulty units produce the opposite answers (one correct and the other incorrect) the overall result will be correct. Assuming that the probability of such a compensating error is c, derive the reliability expression for the TMR configuration.

2. See Ramamoorthy and Han [RAMA 1975]. In order to use parallel redundancy in digital logic, we have to associate an online detector with each unit giving us detector–redundant systems. However, a detector may itself fail. Compare the reliability of a three-unit detector–redundant system with a TMR system (without online detectors). Assume the reliability of a simplex unit is r, the reliability of a detector is d and the reliability of a voter is v. A detector redundant system is said to have failed when all unit–detector pairs have failed and a unit–detector pair is a series combination of the unit and its associated detector.

3. In manufacturing a certain component, two types of defects are likely to occur with respective probabilities 0.05 and 0.1. What is the probability that a randomly chosen component

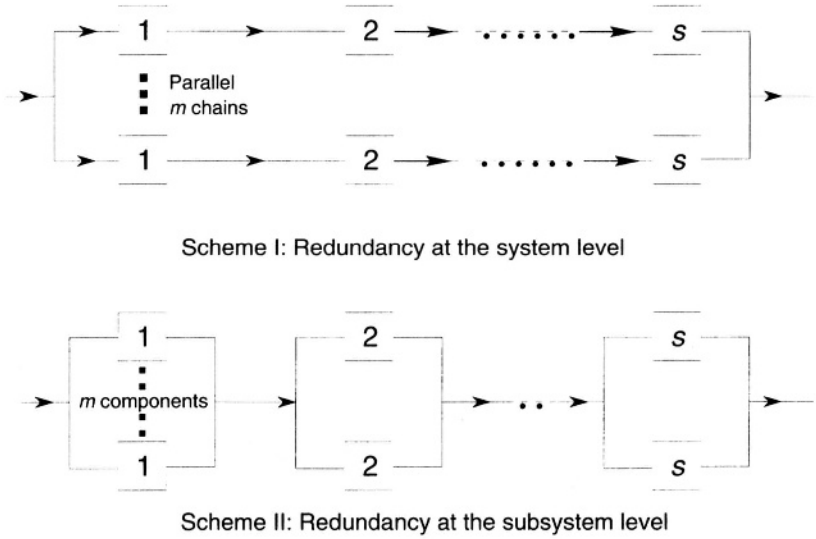
    (a) does not have both kinds of defects?

    (b) is defective?

    (c) has only one kind of defect given that it is found to be defective?

4. Assume that the probability of successful transmission of a single bit over a binary communication channel is p. We desire to transmit a 4-bit word over the channel. To increase the probability of successful word transmission, we may use 7-bit Hamming code (4 data bits + 3 check bits). Such a code is known to be able to correct single-bit errors [HAMM 1980]. Derive the probabilities of successful word transmission under the two schemes and derive the condition under which the use of Hamming code will improve performance.

5. We want to compare two different schemes of increasing reliability of a system using redundancy. Suppose that the system needs s identical components in series for proper operation. Further suppose that we are given $m \cdot s$ components. Out of the two schemes shown in Figure 1.P.7, which one will provide a higher reliability? Given that the reliability of an individual component is r, derive the expressions for the reliabilities of two configurations. For $m = 3$ and $s = 2$, compare the two expressions.



Scheme I: Redundancy at the system level

Scheme II: Redundancy at the subsystem level

Figure 1.P.7. Comparison of two redundancy schemes

6. In three boxes there are capacitors as shown in the following table:

Capacitance	Number in box		
(in μF)	1	2	3
1.0	10	90	25
0.1	50	30	80
0.01	70	90	120

An experiment consists of first randomly selecting a box (assume that each box has the same probability of selection) and then randomly selecting a capacitor from the chosen box.

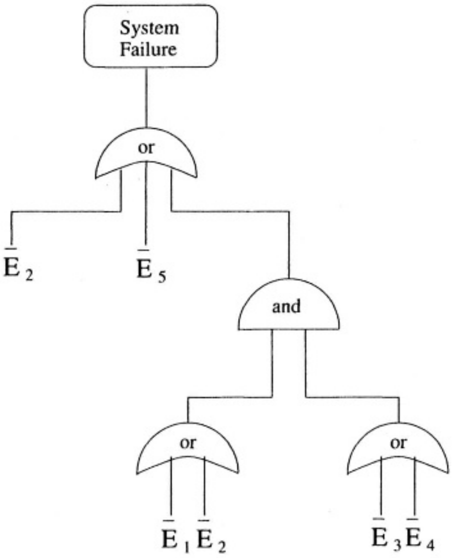

System Failure
or
$\overline{E}_2$
$\overline{E}_5$
and
or
or
$\overline{E}_1$ $\overline{E}_2$
$\overline{E}_3$ $\overline{E}_4$

Figure 1.P.8. A fault tree

(a) What is the probability of selecting a 0.1 μF capacitor, given that box 3 is chosen?

(b) If a 0.1 μF capacitor is chosen, what is the probability that it came from box 1?

(c) List all nine conditional probabilities of capacitor selections, given certain box selections.

7. For the fault tree shown in Figure 1.P.8

   (1) Write down the structure function.

   (2) Derive reliability expressions by

      (a) State enumeration method

      (b) Method of inclusion–exclusion

      (c) Sum of disjoint products method

      (d) Conditioning on the shared event $\overline{E}_2$

8. For the BTS sector/transmitter of Example 1.21, draw the equivalent fault tree, and derive reliability expressions by means of state enumeration, inclusion–exclusion, and SDP methods.

REFERENCES

[ASH 1970] R. B. Ash, *Basic Probability Theory*, J. Wiley, New York, 1970.

[BALA 1996] M. Balakrishnan and K. S. Trivedi, "Stochastic Petri nets for the reliability analysis of communication network applications with alternate-routing," *Reliability Eng. Syst. Safety*, **52**, 243–259 (1996).

[DOSS 2000] K. Doss, personal communication, 2000.

[GOOD 1977] S. E. Goodman and S. Hedetniemi, *Introduction to the Design and Analysis of Algorithms*, McGraw-Hill, New York, 1977.

[HAMM 1980] R. W. Hamming, *Coding and Information Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[HECH 1976] H. Hecht, "Fault-tolerant software for real-time applications," *ACM Comput. Surv.*, 391–408 (Dec. 1976).

[HENL 1981] E. Henley and H. Kumamoto, *Reliability Engineering and Risk Assessment*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[IBE 1989] O. Ibe, R. Howe, and K. S. Trivedi, "Approximate availability analysis of VAXCluster systems," *IEEE Trans. Reliability*, **38**(1), 146–152 (Apr. 1989).

[KOVA 2000] I. Kovalenko, personal communication, 2000.

[LIU 1968] C. L. Liu, *Introduction to Combinatorial Mathematics*, McGraw-Hill, New York, 1968.

[LUO 1998] T. Luo and K. S. Trivedi, "An improved algorithm for coherent system reliability," *IEEE Trans. Reliability*, **47**(1), 73–78 (March 1998).

[MISR 1992] K. B. Misra, *Reliability Analysis and Prediction: A Methodology Oriented Treatment*, Elsevier, Amsterdam, 1992.

[RAI 1995] S. Rai, M. Veeraraghavan, and K. S. Trivedi, "A survey on efficient computation of reliability using disjoint products approach," *Networks*, **25**(3), 147–163 (1995).

[RAMA 1975] C. V. Ramamoorthy and Y.-W. Han, "Reliability analysis of systems with concurrent error detection," *IEEE Trans. Comput.*, 868–878 (Sept. 1975).

[RAND 1978] B. Randell, P. A. Lee, and P. C. Treleaven , "Reliability issues in computing system design," *ACM Comput. Surv.*, **10**(2), 123–166 (June 1978).

[SAHN 1996] R. A. Sahner, K. S. Trivedi, and A. Puliafito, *Performance and Reliability Analysis of Computer System: An Example-Based Approach Using the SHARPE Software Package*, Kluwer Academic Publishers, Boston, 1996.

[STAR 1979] H. Stark and F. B. Tuteur, *Modern Electrical Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[SUN 1999] H.-R. Sun, Y. Cao, K. S. Trivedi, and J. J. Han, "Availability and performance evaluation for automatic protection switching in TDMA wireless system," *Pacific Rim Int. Symp. Dependable Computing, Hong Kong* (*PRDC99*), Dec. 1999, pp. 15–22.

[ZANG 1999] X. Zang, H.-R. Sun, and K. S. Trivedi, "A BDD approach to dependability analysis of distributed computer systems with imperfect coverage," in *Dependable Network Computing,* D. R. Avresky (ed.), Kluwer Academic Publishers, Amsterdam, Dec. 1999, pp. 167–190.