Models for Probability and Statistical Inference

Theory and Applications

JAMES H. STAPLETON

Michigan State University
Department of Statistics and Probability
East Lansing, Michigan

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Models for Probability and
Statistical Inference

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

Models for Probability and Statistical Inference

Theory and Applications

JAMES H. STAPLETON

Michigan State University
Department of Statistics and Probability
East Lansing, Michigan

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Stapleton, James H., 1931–
Models for probability and statistical inference: theory and applications/James H. Stapleton.
p. cm.
ISBN 978-0-470-07372-8 (cloth)
1. Probabilities—Mathematical models. 2. Probabilities—Industrial applications. I. Title.
QA273.S7415 2008
519.2—dc22 2007013726

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Alicia, who has made my first home so pleasant
for almost 44 years.
To Michigan State University and its Department of Statistics
and Probability, my second home for almost 49 years,
to which I will always be grateful.

Contents

Preface

This book was written over a five to six-year period to serve as a text for the two-semester sequence on probability and statistical inference, STT 861–2, at Michigan State University. These courses are offered for master's degree students in statistics at the beginning of their study, although only one-half of the students are working for that degree. All students have completed a minimum of two semesters of calculus and one course in linear algebra, although students are encouraged to take a course in analysis so that they have a good understanding of limits. A few exceptional undergraduates have taken the sequence. The goal of the courses, and therefore of the book, is to produce students who have a fundamental understanding of statistical inference. Such students usually follow these courses with specialized courses on sampling, linear models, design of experiments, statistical computing, multivariate analysis, and time series analysis.

For the entire book, simulations and graphs, produced by the statistical package S-Plus, are included to build the intuition of students. For example, Section 1.1 begins with a list of the results of 400 consecutive rolls of a die. Instructors are encouraged to use either S-Plus or R for their courses. Methods for the computer simulation of observations from specified distributions are discussed.

Each section is followed by a selection of problems, from simple to more complex. Answers are provided for many of the problems.

Almost all statements are backed up with proofs, with the exception of the continuity theorem for moment generating functions, and asymptotic theory for logistic and log-linear models. Simulations are provided to show that the asymptotic theory provides good approximations.

The first six chapters are concerned with probability, the last seven with statistical inference. If a few topics covered in the first six chapters were to be omitted, there would be enough time in the first semester to cover at least the first few sections of Chapter Seven, on estimation. There is a bit too much material included on statistical inference for one semester, so that an instructor will need to make judicious choices of sections. For example, this instructor has omitted Section 7.8, on Fisher information, the Cramér–Rao bound, and asymptotic normality of MLEs, perhaps the most difficult material in the book. Section 7.9, on sufficiency, could be omitted.

Chapter One is concerned with discrete models and random variables. In Chapter Two we discuss discrete distributions that are important enough to have names: the binomial, hypergeometric, geometric, negative binomial, and Poisson, and the Poisson process is described. In Chapter Three we introduce continuous distributions, expected values, variances, transformation, and joint densities.

Chapter Four concerns the normal and gamma distributions. The beta distribution is introduced in Problem 4.3.5. Chapter Five, devoted to conditional distributions, could be omitted without much negative effect on statistical inference. Markov chains are discussed briefly in Chapter Five. Chapter Six, on limit theory, is usually the most difficult for students. Modes of convergence of sequences of random variables, with special attention to convergence in distribution, particularly the central limit theorem for independent random variables, are discussed thoroughly.

Statistical inference begins in Chapter Seven with point estimation: first methods of evaluating estimators, then methods of finding estimators: the method of moments and maximum likelihood. The topics of consistency and the δ-method are usually a bit more difficult for students because they are often still struggling with limit arguments. Section 7.7, on confidence intervals, is one of the most important topics of the last seven chapters and deserves extra time. The author often asks students to explain the meaning of confidence intervals so that "your mother [or father] would understand." Students usually fail to produce an adequate explanation the first time. As stated earlier, Section 7.8 is the most difficult and might be omitted. The same could be said for Section 7.9, on sufficiency, although the beauty of the subject should cause instructors to think twice before doing that.

Chapter Eight, on testing hypotheses, is clearly one of the most important chapters. We hope that sufficient time will be devoted to it to "master" the material, since the remaining chapters rely heavily on an understanding of these ideas and those of Section 7.7, on confidence intervals.

Chapter Nine is organized around the distributions defined in terms of the normal: multivariate normal, chi-square, t, and F (central and noncentral). The usefulness of each of the latter three distributions is shown immediately by the development of confidence intervals and testing methods for "normal models." Some of "Student's" data from the 1908 paper introducing the t-distribution is used to illustrate the methodology.

Chapter Ten contains descriptions of the two- and one-sample Wilcoxon tests, together with methods of estimation based on these. The Kolmogorov–Smirnov one- and two-sample tests are also discussed.

Chapter Eleven, on linear models, takes the linear space-projection approach. The geometric intuition it provides for multiple regression and the analysis of variance, by which sums of squares are simply squared lengths of vectors, is quite valuable. Examples of S-Plus and SAS printouts are provided.

Chapter Twelve begins with logistic regression. Although the distribution theory is quite different than the linear model theory discussed in Chapter Eleven and is asymptotic, the intuition provided by the vector-space approach carries over to logistic regression. Proofs are omitted in general in the interests of time and the students' level

of understanding. Two-way frequency tables are discussed for models which suppose that the logs of expected frequencies satisfy a linear model.

Finally, Chapter Thirteen has sections on survival analysis, including the Kaplan–Meier estimator of the cumulative distribution function, bootstrapping, Bayesian statistics, and sampling. Each is quite brief. Instructors will probably wish to select from among these four topics.

I thank the many excellent students in my Statistics 861–2 classes over the last seven years, who provided many corrections to the manuscript as it was being developed. They have been very patient.

JIM STAPLETON

March 7, 2007

CHAPTER ONE

Discrete Probability Models

1.1 INTRODUCTION

The mathematical study of probability can be traced to the seventeenth-century correspondence between Blaise Pascal and Pierre de Fermat, French mathematicians of lasting fame. Chevalier de Mere had posed questions to Pascal concerning gambling, which led to Pascal's correspondence with Fermat. One question was this: Is a gambler equally likely to succeed in the two games: (1) at least one 6 in four throws of one six-sided die, and (2) at least one double-6 (6–6) in 24 throws of two six-sided dice? At that time it seemed to many that the answer was yes. Some believe that de Mere had empirical evidence that the first event was more likely to occur than the second, although we should be skeptical of that, since the probabilities turn out to be 0.5178 and 0.4914, quite close. After students have studied Chapter One they should be able to verify these, then, after Chapter Six, be able to determine how many times de Mere would have to play these games in order to distinguish between the probabilities.

In the eighteenth century, probability theory was applied to astronomy and to the study of errors of measurement in general. In the nineteenth and twentieth centuries, applications were extended to biology, the social sciences, medicine, engineering—to almost every discipline. Applications to genetics, for example, continue to grow rapidly, as probabilistic models are developed to handle the masses of data being collected. Large banks, credit companies, and insurance and marketing firms are all using probability and statistics to help them determine operating rules.

We begin with discrete probability theory, for which the events of interest often concern count data. Although many of the examples used to illustrate the theory involve gambling games, students should remember that the theory and methods are applicable to many disciplines.

Models for Probability and Statistical Inference: Theory and Applications, By James H. Stapleton

1.2 SAMPLE SPACES, EVENTS, AND PROBABILITY MEASURES

We begin our study of probability by considering the results of 400 consecutive throws of a *fair die*, a six-sided cube for which each of the numbers 1, 2, ..., 6 is equally likely to be the number showing when the die is thrown.

61635	52244	21641	36536	52114	64452	33132	26324	62624	63134
36426	33552	65554	64623	56111	32256	36435	64146	53514	56364
52624	12534	15362	65261	43445	13223	66126	53623	63265	21564
21524	13552	65253	21225	42234	32361	62454	54561	15125	36555
45215	66442	42635	52522	13242	15434	16336	63241	13111	54343
32261	63155	55235	13611	54346	56323	41666	31221	53233	52414
53366	62336	11265	55136	56524	64215	44221	14222	15145	31662
55241	54223	25156	56155	43324	36566	23466	51123	11414	24653

The frequencies are:

1	2	3	4	5	6
60	73	65	58	74	70

We use these data to motivate the definitions and theory to be presented. Consider, for example, the following question: What is the probability that the five numbers appearing in five throws of a die are all different? Among the 80 consecutive sequences of five numbers above, in only four cases were all five numbers different, a relative frequency of $5/80 = 0.0625$. In another experiment, with 2000 sequences of five throws each, all were different 183 times, a relative frequency of 0.0915. Is there a way to determine the long-run relative frequency? Put another way, what could we expect the relative frequency to be in 1 million throws of five dice?

It should seem reasonable that all possible sequences of five consecutive integers from 1 to 6 are equally likely. For example, prior to the 400-throw experiment, each of the first two sequences, 61635 and 52244, were equally likely. For this example, such five-digit sequences will be called *outcomes* or *sample points*. The collection of all possible such five-digit sequences will be denoted by S, the sample space. In more mathematical language, S is the Cartesian product of the set $A = \{1, 2, 3, 4, 5, 6\}$ with itself five times. This collection of sequences is often written as $A^{(5)}$. Thus, $S = A^{(5)} = A \times A \times A \times A \times A$. The number of outcomes (or sample points) in S is $6^5 = 7776$. It should seem reasonable to suppose that all outcomes (five-digit sequences) have probability $1/6^5$.

We have already defined a *probability model* for this experiment. As we will see, it is enough in cases in which the sample space is discrete (finite or countably infinite) to assign probabilities, nonnegative numbers summing to 1, to each outcome in the sample space S. A discrete probability model has been defined for an experiment when (1) a finite or countably infinite sample space has been defined, with each possible result of the experiment corresponding to exactly one outcome; and (2) probabilities, nonnegative numbers, have been assigned to the outcomes in such a way that they

sum to 1. It is not necessary that the probabilities assigned all be the same as they are for this example, although that is often realistic and convenient.

We are interested in the *event* A that all five digits in an outcome are different. Notice that this event A is a subset of the sample space S. We say that an event A has *occurred* if the outcome is a member of A. In this case event A did not occur for any of the eight outcomes in the first row above.

We define the *probability* of the event A, denoted $P(A)$, to be the sum of the probabilities of the outcomes in A. By defining the probability of an event in this way, we assure that the probability measure P, defined for all subsets (events, in probability language) of S, obeys certain axioms for probability measures (to be stated later). Because our probability measure P has assigned all probabilities of outcomes to be equally likely, to find $P(A)$ it is enough for us to determine the number of outcomes $N(A)$ in A, for then $P(A) = N(A)[1/N(S)] = N(A)/N(S)$. Of course, this is the case only because we assigned equal probabilities to all outcomes.

To determine $N(A)$, we can apply the multiplication principle. A is the collection of 5-tuples with all components different. Each outcome in A corresponds to a way of filling in the boxes of the following cells:

The first cell can hold any of the six numbers. Given the number in the first cell, and given that the outcome must be in A, the second cell can be any of five numbers, all different from the number in the first cell. Similarly, given the numbers in the first two cells, the third cell can contain any of four different numbers. Continuing in this way, we find that $N(A) = (6)(5)(4)(3)(2) = 720$ and that $P(A) = 720/7776 = 0.0926$, close to the value obtained for 2000 experiments. The number $N(A) = 720$ is the number of permutations of six things taken five at a time, indicated by $P(6, 5)$.

Example 1.2.1 Consider the following discrete probability model, with sample space $S = \{a, b, c, d, e, f\}$.

Outcome ω	a	b	c	d	e	f
$P(\omega)$	0.30	0.20	0.25	0.10	0.10	0.05

Let $A = \{a, b, d\}$ and $B = \{b, d, e\}$. Then $A \cup B = \{a, b, d, e\}$ and $P(A \cup B) = 0.3 + 0.2 + 0.1 + 0.1 = 0.7$. In addition, $A \cap B = \{b, d\}$, so that $P(A \cap B) = 0.2 + 0.1 = 0.3$. Notice that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. (Why must this be true?). The *complement* of an event D, denoted by D^c, is the collection of outcomes in S that are not in D. Thus, $P(A^c) = P(\{c, e, f\}) = 0.15 + 0.15 + 0.10 = 0.40$. Notice that $P(A^c) = 1 - P(A)$. Why must this be true? □

Let us consider one more example before more formally stating the definitions we have already introduced.

Example 1.2.2 A penny and a dime are tossed. We are to observe the number X of heads that occur and determine $P(X = k)$ for $k = 0, 1, 2$. The symbol X, used here for its convenience in defining the events $[X = 0]$, $[X = 1]$, and $[X = 2]$, will be called a *random variable* (rv). $P(X = k)$ is shorthand for $P([X = k])$. We delay a more formal discussion of random variables.

Let $S_1 = \{\text{HH, HT, TH, TT}\} = \{\text{H, T}\}^{(2)}$, where the result for the penny and dime are indicated in this order, with H denoting head and T denoting tail. It should seem reasonable to assign equal probabilities $1/4$ to each of the four outcomes. Denote the resulting probability measure by P_1. Thus, for $A =$ [event that the coins give the same result] $= \{\text{HH, TT}\}$, $P_1(A) = 1/4 + 1/4 = 1/2$. □

The 400 throws of a die can be used to simulate 400 throws of a coin, and therefore 200 throws of two coins, by considering 1, 2, and 3 as heads and 4, 5, and 6 as tails. For example, using the first 10 throws, proceeding across the first row, we get TH, TH, TT, HH, TT. For all 400 die throws, we get 50 cases of HH, 55 of HT, 47 of TH, and 48 of TT, with corresponding relative proportions 0.250, 0.275, 0.235, and 0.240. For the experiment with 10,000 throws, simulating 5000 pairs of coin tosses, we obtain 1288 HH's, 1215 HT's, 1232 TH's, and 1265 TT's, with relative frequencies 0.2576, 0.2430, 0.2464, and 0.2530. Our model (S_1, P_1) seems to fit well.

For this model we get $P_1(X = 0) = 1/4$, $P_1(X = 1) = 1/4 + 1/4 = 1/2$, and $P_1(X = 2) = 1/4$. If we are interested only in X, we might consider a slightly smaller model, with sample space $S_2 = \{0, 1, 2\}$, where these three outcomes represent the numbers of heads occurring. Although it is tempting to make the model simpler by assigning equal probabilities $1/3, 1/3, 1/3$ to these outcomes, it should be obvious that the empirical results of our experiments with 400 and 10,000 tosses are not consistent with such a model. It should seem reasonable, instead, to assign probabilities $1/4, 1/2, 1/4$, thus defining a probability measure P_2 on S_2. The model (S_2, P_2) is a *recoding* or *reduction* of the model (S_1, P_1), with the outcomes HT and TH of S_1 corresponding to the single outcome $X = 1$ of S_2, with corresponding probability determined by adding the probabilities $1/4$ and $1/4$ of HT and TH.

The model (S_2, P_2) is simpler than the model (S_1, P_1) in the sense that it has fewer outcomes. On the other hand, it is more complex in the sense that the probabilities are unequal. In choosing appropriate probability models, we often have two or more possible models. The choice of a model will depend on its approximation of experimental evidence, consistency with fundamental principles, and mathematical convenience.

Let us stop now to define more formally some of the terms already introduced.

Definition 1.2.1 A *sample space* is a collection S of all possible results, called *outcomes*, of an experiment. Each possible result of the experiment must correspond to one and only one outcome in S. A sample space is *discrete* if it has a finite or countably infinite number of outcomes. (A set is *countably infinite* if it can be put into one-to-one correspondence with the positive integers.) □

Definition 1.2.2 An *event* is a subset of a sample space. An event A is said to *occur* if the outcome of an experiment is a member of A. □

Definition 1.2.3 A *probability measure* P on a discrete sample space S is a function defined on the subsets of S such that:

(a) $P(\{\omega\}) \geq 0$ for all points $\omega \in S$.
(b) $P(A) = \sum_{\omega \in A} P(\omega)$ for all subsets A of S.
(c) $P(S) = 1$.

For simplicity, we write $P(\{\omega\})$ as $P(\omega)$. □

Definition 1.2.4 A *probability model* is a pair (S, P), where P is a probability measure on S. □

In writing $P(\{\omega\})$ as $P(\omega)$, we are abusing notation slightly by using the symbol P to denote both a function on S and a function on the subsets of S. We assume that students are familiar with the notation of set theory: *union*, $A \cup B$; *intersection* $A \cap B$; and *complement*, A^c. Thus, for events A and B, the event $A \cup B$ is said to occur if the outcome is a member of A or B (by "or" we include the case that the outcome is in both A and B). The event $A \cap B$ is said to occur if both A and B occur. A^c, called a complement, is said to occur if A does not occur. For convenience we sometimes write $A \cap B$ as AB.

We also assume that the student is familiar with the notation for relationships among sets, $A \subset B$ and $A \supset B$. Thus, if $A \subset B$, the occurrence of event A implies that B must occur. We sometimes use the language "event A implies event B." For the preceding two-coin-toss example, the event $[X = 1]$ implies the event $[X \geq 1]$.

Let $\emptyset$ denote the *empty event*, the subset of S consisting of no outcomes. Thus, $A \cap A^c = \emptyset$. We say that two events A and B are *mutually exclusive* if their intersection is empty. That is, $A \cap B = \emptyset$. Thus, if A and B are mutually exclusive, the occurrence of one of them implies that the other cannot occur. In set-theoretic language we say that A and B are *disjoint*. *DeMorgan's laws* give relationships among intersection, union, and complement:

$$(1) \quad (A \cap B)^c = A^c \cup B^c \quad \text{and} \quad (2) \quad (A \cup B)^c = A^c \cap B^c.$$

These can be verified from a *Venn diagram* or by showing that any element in the set on the left is a member of the set on the right, and vice versa (see Figure 1.2.1).

Properties of a Probability Measure P on a Sample Space S

1. $P(\emptyset) = 0$.
2. $P(S) = 1$.
3. For any event A, $P(A^c) = 1 - P(A)$.

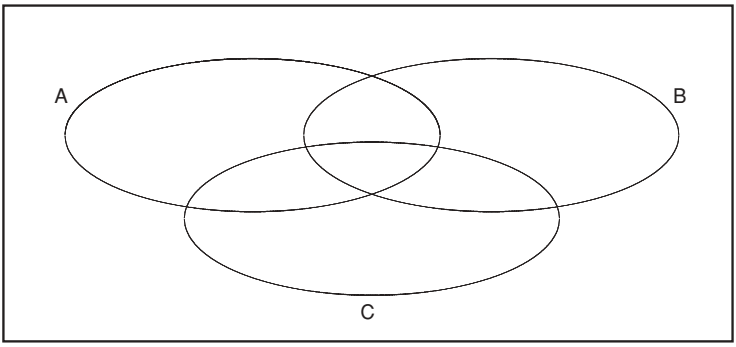

FIGURE 1.2.1 Venn diagram for three events.

4. For any events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. For three events A, B, C, $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$. This follows from repeated use of the identity for two events. An almost obvious similar result holds for the probability of the union of n events, with $2^n - 1$ terms on the right.
5. For events A and B with $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$. More generally, if $A_1, A_2, \ldots,$ are disjoint (mutually exclusive) events, $P(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$. This property of P is called *countable additivity*. Since A_k for $k > n$ could be $\emptyset$, the same equality holds when ∞ is replaced by any integer $n > 0$.

Let us make use of some of these properties in a few examples.

Example 1.2.3 Smith and Jones each throw three coins. Let X denote the number of heads for Smith. Let Y denote the number of heads for Jones. Find $P(X = Y)$.

We can simulate this experiment using the 400-die-tossing example, again letting 1, 2, 3 correspond to heads, 4, 5, 6 correspond to tails. Let the first three tosses be for Smith, the next three for Jones, so that the first six tosses determine one trial of the experiment. Repeating, going across rows, we get 36 trials of the experiment. Among these 36 trials, 10 resulted in $X = Y$, suggesting that $P(X = Y)$ may be approximately $10/36 = 0.278$. For the experiment with 9996 tosses, 525 among 1666 trials gave $X = Y$, suggesting that $P(X = Y)$ is close to $525/1666 = 0.3151$. Let us now try to find the probability by mathematical methods.

Let $S_1 = \{\text{H, T}\}^{(3)}$, the collection of 3-tuples of heads and tails. S_1 is the collection of outcomes for Smith. Also, let $S_2 = \{\text{H, T}\}^{(3)} = S_1$, the collection of outcomes for Jones. Let $S = S_1 \times S_2$. This Cartesian product can serve as the sample space for the experiment in which Smith and Jones both toss three coins. One outcome in S, for example (using shorthand notation), is (HTH, TTH), so that $X = 2$, $Y = 1$. The event $[X = Y]$ did not occur. Since $N(S_1) = N(S_2) = 2^3 = 8$, $N(S) = 64$. Define the probability measure P on S by assigning probability $1/64$ to each outcome. The pair (S, P) constitutes a probability model for the experiment.

Let $A_k = [X = Y = k]$ for $k = 0, 1, 2, 3$. By this bracket notation we mean the collection of outcomes in S for which X and Y are both k. We might also have

TABLE 1.2.1 Box Diagram

	R_2	R_2^c	
R_1		0.2	0.6
R_1^c			
	0.5		1.0

written $A_k = [X = k, Y = k]$. The events A_0, A_1, A_2, A_3 are mutually exclusive, and $[X = Y] = A_0 \cup A_1 \cup A_2 \cup A_3$. It follows from property 5 above that $P(X = Y) = P(A_0) + P(A_1) + P(A_2) + P(A_3)$. Since $N(A_0) = 1$, $N(A_1) = 3^2$, $N(A_2) = 3^2$, $N(A_3) = 1$, and $P(A_k) = N(A_k)/64$, we find that $P(X = Y) = 20/64 = 5/16 = 0.3125$, relatively close to the proportions obtained by experimentation. □

Example 1.2.4 Suppose that a probability model for the weather for two days has been defined in such a way that $R_1 = [\text{rain on day 1}]$, $R_2 = [\text{rain on day 2}]$, $P(R_1) = 0.6$, $P(R_2) = 0.5$, and $P(R_1 R_2^c) = 0.2$. Find $P(R_1 R_2)$, $P(R_1^c R_2)$, and $P(R_1 \cup R_2)$.

Although a Venn diagram can be used, a *box diagram* (Table 1.2.1) makes things clearer. From the three probabilities given, the other cells may be determined by subtraction. Thus, $P(R_1^c) = 0.4$, $P(R_2^c) = 0.5$, $P(R_1 R_2) = 0.4$, $P(R_1^c R_2) = 0.1$, $P(R_1^c R_2^c) = 0.3$, $P(R_1^c \cup R_2^c) = 0.6$. Similar tables can be constructed for the three events. □

Example 1.2.5 A jury of six is to be chosen randomly from a panel of eight men and seven women. Let X denote the number of women chosen. Let us find $P(X = k)$ for $k = 0, 1, \ldots, 6$.

For convenience, name the members of the panel $1, 2, \ldots, 15$, with the first eight being men. Let $D = \{1, 2, \ldots, 15\}$. Since the events in which we are interested do not depend on the order in which the people are drawn, the outcomes can be chosen to be subsets of D of size 6. That is, $S = \{B \mid B \subset D, N(B) = 6\}$. We interpret "randomly" to mean that all the outcomes in S should have equal probability. We need to determine $N(S)$. Such subsets are often called *combinations*. The number of combinations of size k of a set of size n is denoted by $\binom{n}{k}$. Thus, $N(S) = \binom{15}{6}$.

The number of *permutations* (6-tuples of different people) of 15 people six at a time, is $P(15, 6) = (15)(14)(13)(12)(11)(10) = 15!/9!$. The number of ways of ordering six people is $P(6, 6) = 6!$. Since (number of subsets of D of size 6) $\times$ (number of ways of ordering six people) $= P(15, 6)$, we find that $N(S) = \binom{15}{6} = P(15, 6)/6! = 15!/[9!6!] = 5005$. Each outcome is assigned probability $1/5005$.

Consider the event $[X = 2]$. An outcome in $[X = 2]$ must include exactly two females and therefore four males. There are $\binom{7}{2} = (7)(6)/(2)(1) = 21$ such combinations. There are $\binom{8}{4}$ combinations of four males. There are

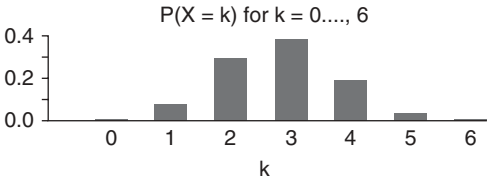

FIGURE 1.2.2 Probability mass function for X.

therefore $\binom{7}{2}\binom{8}{4} = (21)(70) = 1470$ outcomes in the event $[X = 2]$. Therefore, $P(X = 2) = \binom{7}{2}\binom{8}{4}\Big/\binom{15}{6} = 1470/5005 = 0.2937$.

Similarly, we find $P(X = 3) = \binom{7}{3}\binom{8}{3}/N(S) = (35)(56)/5005 = 1960/5005 = 0.3916$, $P(X = 0) = 28/5005 = 0.0056$, $P(X = 1) = 392/5005 = 0.0783$, $P(X = 4) = 980/5005 = 0.1691$, $P(X = 5) = 168/5005 = 0.0336$, $P(X = 6) = 7/5005 = 0.0014$. Figure 1.2.2 shows that these probabilities go "uphill," then "downhill," with the maximum at 3. □

Example 1.2.6 (Poker) The cards in a 52-card deck are classified in two ways: by 13 ranks and by four suits. The 13 ranks and four suits are indicated by the column and row headings in Table 1.2.2. In the game of poker, five cards are chosen randomly without replacement, so that all possible subsets (called *hands*) are equally likely. Hands are classified as follows, with decreasing worth: straight flush, 4-of-a-kind, full house, flush, straight, 3-of-a-kind, two pairs, one pair, and "bad." The category "bad" was chosen by the author so that all hands fall in one of the categories. *k-of-a-kind* means that k cards are of one rank but the other $5 - k$ cards are of differing ranks. A straight consists of five cards with five consecutive ranks. For this purpose the ace is counted as either high or low, so that ace−2−3−4−5 and 10−J−Q−K−ace both constitute straights. A flush consists of cards that are all of the same suit. So that a hand falls in exactly one of these categories, it is always classified in the higher category if it satisfies the definition. Thus, a hand that is both a straight and a flush is classified as a straight flush but not as a straight or as a flush. A full house has three cards of one rank, two of another. Such a hand is not counted as 3-of-a-kind or as 2-of-a-kind.

TABLE 1.2.2 52-Card Deck

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Spades													
Hearts													
Diamonds													
Clubs													

Let D be the set of 52 cards. Let S be the collection of five-card hands, subsets of five cards. Thus, $S = \{B \mid B \subset D, N(B) = 5\}$. "Randomly without replacement" means that all $N(S) = \binom{52}{5} = 2{,}598{,}960$ outcomes are equally likely. Thus, we have defined a probability model.

Let F be the event [full house]. The rank having three cards can be chosen in 13 ways. For each of these the rank having two cards can be chosen in 12 ways. For each choice of the two ranks there are $\binom{4}{3}\binom{4}{2} = (4)(6)$ choices for the suits. Thus, $N(F) = (13)(12)(6)(4) = 3744$, and $P(F) = 3744/2,598,960 = 0.0001439 = 1/694$. Similarly, $P(\text{straight}) = 10[4^5 - 4]/N(S) = 10{,}200/N(S) = 0.003925 = 1/255$, and $P(2\,\text{pairs}) = \binom{13}{2}\binom{4}{2}\binom{4}{2}(44)/N(S) = 123{,}552/N(S) = 0.04754 = 1/21.035$. In general, as the value of a hand increases, the probability of the corresponding category decreases (see Problem 1.2.3). □

Example 1.2.7 (The Birthday Problem) A class has n students. What is the probability that at least one pair of students have the same birthday, not necessarily the same birth year?

So that we can think a bit more clearly about the problem, let the days be numbered $1, \ldots, 365$, and suppose that $n = 20$. Birth dates were randomly chosen using the function "sample" in S-Plus, a statistical computer language.

(1)	52	283	327	15	110	214	141*	276	16	43	130	219	337	234	64	262	141*	336	220	10
(2)	331	106	364	219	209	70	11	54	192	360	75	228	132	172	30	5	166	15	143	173
(3)	199	361*	211	48	86	129	39	202	339	347	22	361*	208	276	75	115	65	291	57	318
(4)	300	252	274	135	118	199	254	316	133	192	238	189	94	167	182	5	235	363	160	214
(5)	110	187	107	47	250	341	49	341	258	273	290	225	31	108	334	118	214	87	315	282
(6)	195	270^	24	204#	69	233	38%	204#	12*	358	38%	138	149	76	71	186	106	270^	12*	87
(7)	105	354	259	10	244	22	70	28	278	127	320	238	60	8	165	339	119	346	295	92
(8)	359#	289	112	299	201	36	94	75	269	359#	122	288	310	329	133	117	291	61*	61*	336
(9)	300	346	72	296	221	176	109	189	3	114	83	222	292	318	238	215	246	183	220	236
(10)	337	98	17	357	75	32	138	255	150	12	88	133	135	5	319	198	119	288	183	359

Duplicates are indicated by *'s, #'s, ^'s, and %'s. Notice that these 10 trials had 1, 0, 0, 3, 0, 2, 0, 0 duplicates. Based on these trials, we estimate the probability of at least one duplicate to be 4/10. This would seem to be a good estimate, since 2000 trials produced 846 cases with at least one duplicate, producing the estimate 0.423. Let us determine the probability mathematically.

Notice the similarity of this example to the die-throw example at the beginning of the chapter. In this case let $D = \{1, \ldots, 365\}$, the "dates" of the year. Let $S = D^{(n)}$, the n-fold Cartesian product of D with itself. Assign probability $1/N(S) = 1/365^n$ to each outcome. We now have a probability model. □

Let A be the event of at least one duplicate. As with most "at least one" events, it is easier to determine $N(A^c)$ than $N(A)$ directly. In fact, $N(A^c) = P(365, n) = 365(364)\cdots(365 - n + 1)$. Let $G(n) = P(A^c)$. It follows that $G(n) = N(A^c)/N(S) = \prod_{k=1}^{n}[(365 - k + 1)/365] = \prod_{k=1}^{n}[1 - (k - 1)/365]$. We can find

TABLE 1.2.3 Probabilities of Coincident Birthdays

	n								
	10	20	22	23	30	40	50	60	70
$P(A)$	0.1169	0.4114	0.4757	0.5073	0.7063	0.8912	0.9704	0.9941	0.9991
$h(n)$	0.1160	0.4058	0.4689	0.5000	0.6963	0.8820	0.9651	0.9922	0.9987

a good approximation by taking logarithms and converting the product to a sum. $\ln\ G(n) = \sum_{k=1}^{n} \ln\,[1-(k-1)/365]$. For x close to zero, $\ln(1-x)$ is close to $-x$, the Taylor series linear approximation. It follows that for $(n-1)/365$ small, $\ln(G(n))$ is approximately $-\sum_{k=1}^{n}[(k-1)/365] = -[n(n-1)/2]/365 = -n(n-1)/730$. Hence, a good approximation for $P(A) = 1 - G(n)$ is $h(n) = 1 - e^{-n(n-1)/730}$. Table 1.2.3 compares $P(A)$ to its approximation $h(n)$ for various n. Notice that the relative error in the approximation of $P(A^c)$ by $1 - h(n)$ increases as n increases.

Pascal's Triangle: An Interesting Identity for Combinatorics

Consider a set of five elements $A = \{a_1, a_2, \ldots, a_5\}$. A has $\binom{5}{3} = 10$ subsets of size 3. These are of two types: those that contain element a_1 and those that do not. The number that contains a_1 is the number of subsets $\binom{4}{2} = 6$ of $\{a_2, \ldots, a_5\}$ of size 2. The number of subsets of A that do not contain a_1 is $\binom{4}{3} = 4$. Thus, $\binom{5}{3} = \binom{4}{2} + \binom{4}{3}$.

More generally, if A has n elements $\{a_1, a_2, \ldots, a_n\}$, A has $\binom{n}{k}$ subsets of size k for $0 < k \leq n$. These subsets are of two types, those that contain a_1 and those that do not. It follows by the same reasoning that $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. The same equality can be proved by manipulation of factorials. Pascal, in the mid-seventeenth century, represented this in the famous *Pascal triangle* (Figure 1.2.3). Each row begins and ends with 1, and each interior value is the sum of the two immediately above. The nth row for $n = 0, 1, \ldots$ has $\binom{n}{k}$ in the kth place for $k = 0, 1, \ldots, n$. Row $n = 4$ has elements 1, 4, 6, 4, 1. Notice that these sum to $16 = 2^4$.

The Equality $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^k$

The collection B of subsets of a set with n elements is in one-to-one correspondence to the set $C = \{0, 1\}^{(n)}$. For example, for the set $A = \{a_1, a_2, a_3, a_4\}$, the point (0, 1, 1, 0) in C corresponds to the subset $\{a_2, a_3\}$, and $\{1, 0, 1, 1\}$ corresponds to the subset $\{a_1, a_3, a_4\}$. Thus, $N(B) = N(C) = 2^k$. But we can count the elements in B another way. There are those with no elements, those with one, those with 2, and so on. The

$$\binom{0}{0} = 1$$

$$\binom{1}{0} \qquad \binom{1}{1}$$

$$\binom{2}{0} \qquad \binom{2}{1} \qquad \binom{2}{2}$$

$$\binom{3}{0} \qquad \binom{3}{1} \qquad \binom{3}{2} \qquad \binom{3}{3}$$

.....

FIGURE 1.2.3 Pascal's triangle.

equality above follows. For example, $2^5 = 1 + 5 + 10 + 10 + 5 + 1$. The sum of the numbers in the row of Pascal's triangle labeled n is 2^n.

Relationship Between Two Probability Models

Example 1.2.8 Suppose that a coin is tossed twice. This experiment may be reasonably modeled by (S_1, P_1), where $S_1 = \{\text{HH, HT, TH, TT}\}$ and P_1 assigns probability $1/4$ to each outcome, and by (S_2, P_2), where $\{X = 0, X = 1, X = 2\}$, where $X =$ (no. heads), and P_2 assigns probabilities $1/4, 1/2, 1/4$. In this case we have essentially "glued" together the outcomes HT and TH in S_1 to create one outcome $(X = 1)$ in S_2. We have also added the probabilities $1/4$ and $1/4$ in (S_1, P_1) to get $P_2(X = 1) = 1/2$. The model (S_2, P_2) is simpler than (S_1, P_1) in the sense that it has fewer outcomes, but it is more complex in the sense that the probabilities aren't equal. We will say that the probability model (S_2, P_2) is a *reduction* of the probability model (S_1, P_1), and that (S_1, P_1) is an *expansion* of the probability model (S_2, P_2).

The model (S_1, P_1) can be used to determine the probability of the event $H_1 =$ [first toss is heads]. The probability model (S_2, P_2) cannot be used to determine the probability of H_1. The event H_1 is not "measurable" with respect to the model (S_2, P_2). Such questions on measurability are considered as part of the subject of measure theory. We say very little about it here. □

In order to consider a general definition of reduction and expansion of probability models, we need to recall that for any function $g\colon S_1 \to S_2$,

$$g^{-1}(A) = \{\omega \mid g(\omega) \in A\} \quad \text{for any subset } A \text{ of } S_2.$$

Definition 1.2.5 Let (S_1, P_1) and (S_2, P_2) be two discrete probability models. Then (S_2, P_2) is said to be a *reduction* of (S_1, P_1) if there exists a function g from S_1 to

S_2 such that $P_1(g^{-1}(A)) = P_2(A)$ for any subset A of S_2. (S_1, P_1) is said to be an *expansion* of (S_2, P_2). □

If S_1 is finite or countably infinite, $P_1(g^{-1}(A)) = P_2(A)$ is assured if it holds whenever A is a one-point set. This follows from the fact that $g^{-1}(A) = \{g^{-1}(\omega) \,|\, \omega \in A\}$ is a union of mutually exclusive events.

If X is a discrete random variable defined on (S_1, P_1), let $S_2 = \{k | P(X = k) > 0\}$. Let $P_2(k) = P_1(X = k)$. Then X plays the role of g in the definition so that (S_2, P_2) is a reduction of (S_1, P_1). This is the most common way to reduce a probability model. More generally, if $X_1, \ldots, X_n$ are random variables defined on (S_1, P_1), $\mathbf{X} = (X_1, \ldots, X_n)$, then we can take $S_2 = \{\mathbf{x} \in R_n \,|\, P(\mathbf{X} = \mathbf{x}) > 0\}$ and assign $P_2(\mathbf{x}) = P_1(\mathbf{X} = \mathbf{x})$.

Example 1.2.9 A husband and wife and two other couples are seated at random around a round table with six seats. What is the probability that the husband and wife in a particular couple, say C_1, are seated in adjacent seats?

Let the people be $a, b, \ldots, g$, let the seats be numbered $1, \ldots, 6$, reading clockwise around the table, and let $(x_1, \ldots, x_6)$, where each x_i is one of the these letters, all different, correspond to the outcome in which person x_i is seated in seat i, $i = 1, 2, \ldots, 6$. Let S_1 be the collection of such arrangements. Let P_1 assign probability $1/6!$ to each outcome. Let A be the collection of outcomes for which f and g are adjacent. If f and g are the husband and wife in C_1, then fg in this order may be in seats $12, 23, \ldots, 56, 61$. They may also be in the order of $21, 32, \ldots, 16$. For each of these the other four people may be seated in 4! ways. Thus, $N(A) = (2)(6)(4!) = 288$ and $P(A) = 2/5$.

We may instead let an outcome designate only the seats given to the husband and wife in C_1, and let S_2 be the set of pairs (x, y), $x \neq y$. We have combined all 4! seating arrangements in S_1 which lead to the same seats for the husband and wife in C_1. Thus, $N(S_2) = (6)(5)$. Let P_2 assign equal probability $1/[(5)(6)]$ to each outcome, $4! = 24$ times as large as for the outcomes in S_1. Let $B =$ [husband and wife in C_1 are seated together] $= \{12, 23, \ldots, 61, 21, \ldots, 16\}$, a subset of S_2. Then $P(B) = (2)(6)/(6)(5) = 2/5$, as before. Of course, if we were asked the probability of the event D that all three couples are seated together, each wife next to her husband. we could not answer the question using (S_2, P_2), although we could using the model (S_1, P_1). $P_1(D) = (2)(3!)(2^3)/6! = 96/720 = 2/15$. (Why?) D is an event with respect to S_1 (a subset of S_1), but there is no corresponding subset of S_2. □

Problems for Section 1.2

1.2.1 Consider the sample space $S = \{a, b, c, d, e, f\}$. Let $A = \{a, b, c\}$, $B = \{b, c, d\}$, and $C = \{a, f\}$. For each outcome x in S, let $P(\{x\}) = p(x)$, where $p(a) = 0.20$, $p(b) = 0.15$, $p(c) = 0.20$, $p(d) = 0.10$, $p(e) = 0.30$. Find $p(f)$, $P(A)$, $P(B)$, $P(A \cup B)$, $P(A \cup B^c)$, $P(A \cup B^c \cup C)$. Verify that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2.2 Box 1 has four red and three white balls. Box 2 has three red and two white balls. A ball is drawn randomly from each box. Let (x, y) denote an outcome for which the ball drawn from box 1 has color x and the ball drawn from box 2 has color y. Let $C = \{\text{red, white}\}$ and let $S = C \times C$.

(a) Assign probabilities to the outcomes in S in a reasonable way. [In 1000 trials of this experiment the outcome (red, white) occurred 241 times.]

(b) Let $X =$ (no. red balls drawn). Find $P(X = k)$ for $k = 0, 1, 2$. (In 1000 trials the events $[X = 0]$, $[X = 1]$, and $[X = 2]$ occurred 190, 473, and 337 times.)

1.2.3 For the game of poker, find the probabilities of the events [straight flush], [4-of-a-kind], [flush], [3-of-a-kind], [one pair].

1.2.4 Find the elements in the rows labeled $n = 6$ and $n = 7$ in Pascal's triangle. Verify that their sums are 2^6 and 2^7.

1.2.5 A coin is tossed five times.

(a) Give a probability model so that S is a Cartesian product.

(b) Let $X =$ (no. heads). Determine $P(X = 2)$.

(c) Use the die-toss data at the beginning of the chapter to simulate this experiment and verify that the relative frequency of cases for which the event $[X = 2]$ occurs is close to $P(X = 2)$ for your model.

1.2.6 For a Venn diagram with three events A, B, C, indicate the following events by darkening the corresponding region:

(a) $A \cup B^c \cup C$.

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

(c) $(A \cup B^c) \cap (B^c \cup C)$.

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

1.2.7 Two six-sided fair dice are thrown.

(a) Let $X =$ (total for the two dice). State a reasonable model and determine $P(X = k)$ for $k = 2, 3, \ldots, 12$. (Different reasonable people may have sample spaces with different numbers of outcomes, but their answers should be the same.)

(b) Let $Y =$ (maximum for the two dice). Find $P(Y = j)$ for $j = 1, 2, \ldots, 6$.

1.2.8 **(a)** What is the (approximate) probability that at least two among five nonrelated people celebrate their birthdays in the same month? State a model first. In 100,000 simulations the event occurred 61,547 times.

(b) What is the probability that at least two of five cards chosen randomly without replacement from the deck of 48 cards formed by omitting the aces are of the same rank? Intuitively, should the probability be larger

or smaller than the answer to part (a)? Why? In 100,000 simulations the event occurred 52,572 times.

1.2.9 A small town has six houses on three blocks, $B_1 = \{a, b, c\}$, $B_2 = \{d, e\}$, $B_3 = \{f\}$. A random sample of two houses is to be chosen according to two different methods. Under method 1, all possible pairs of houses are written on slips of paper, the slips are thoroughly mixed, and one slip is chosen. Under method 2, two of the three blocks are chosen randomly without replacement, then one house is chosen randomly from each of the blocks chosen. For each of these two methods state a probability model, then use it to determine the probabilities of the events [house a is chosen], [house d is chosen], [house f is chosen], and [at least one of houses a, d is chosen].

1.2.10 Four married couples attend a dance. For the first dance the partners for the women are randomly assigned among the men. What is the probability that at least one woman must dance with her husband?

1.2.11 From among nine men and seven women a jury of six is chosen randomly. What is the probability that two or fewer of those chosen are men?

1.2.12 A six-sided die is thrown three times.

(a) What is the probability that the numbers appearing are in increasing order? *Hint:* There is a one-to one correspondence between subsets of size 3 and increasing sequences from $\{1, 2, \ldots, 6\}$. In 10,000 simulations the event occurred 934 times.

(b) What is the probability that the three numbers are in nondecreasing order? (2, 4, 4) is not in increasing order, but *is* in nondecreasing order. Use the first 60 throws given at the beginning of the chapter to simulate the experiment 20 times. For the 10,000 simulations, the event occurred 2608 times.

1.2.13 Let (S, P) be a probability model and let A, B, C be three events such that $P(A) = 0.55$, $P(B) = 0.60$, $P(C) = 0.45$, $P(A \cap B) = 0.25$, $P(A \cap C) = 0.20$, $P(B^c \cap C) = 0.15$, and $P(A \cap B \cap C) = 0.10$.

(a) Present a box diagram with $2^3 = 8$ cells giving the probabilities of all events of the form $A^* \cap B^* \cap C^*$, where A^* is either A or A^c, and B^* and C^* are defined similarly.

(b) Draw a Venn diagram indicating the same probabilities.

(c) Find $P(A^c \cap B \cap C^c)$ and $P(A \cup B^c \cup C)$. *Hint:* Use one of DeMorgan's laws for the case of three events.

1.2.14 (*The Matching Problem*)

(a) Let $A_1, \ldots, A_n$ be n events, subsets of the sample space S. Let S_k be the sum of the probabilities of the intersections of all $\binom{n}{k}$ choices of these n events, taken k at a time. For example, for $n = 4$,

$S_3 = P(A_1A_2A_3) + P(A_1A_2A_4) + P(A_1A_3A_4) + P(A_2A_3A_4)$. Prove that $P(A_1 \cup A_2 \cup \cdots \cup A_n) = S_1 - S_2 + \cdots + (-1)^{n+1}S_n$.

(b) Let $X = (X_1, \ldots, X_n)$ be a random permutation of the integers $1, \ldots, n$. Let $A_i = [X_i = i]$. Thus, A_i is the event of a *match* in the ith place. Express the probability of at least one match as a sum of n terms, and then use this to find an approximation for large n. For 1000 simulations with $n = 10$, the frequencies $f(k)$ of k matches were as follows: $f(0) = 351$, $f(1) = 372$, $f(2) = 195$, $f(3) = 60$, $f(4) = 14$, $f(5) = 8$, for an average of 1.038 matches per experiment. The probabilities for the numbers of matches for $n = 3$, are $f(0) = 1/3$, $f(1) = 3/6$, $f(3) = 1/6$. Later we will be able to show that the "expected number" of matches per experiment is 1.0.

(c) Apply the formulas obtained in part (b) to answer Problem 1.2.8.

1.2.15 Give two models (S_i, P_i) for $i = 1, 2$ for the two tosses of a six-sided die, so that (S_2, P_2) is a reduction of (S_1, P_1). Both should enable you to determine the probability that the sum of the numbers appearing exceeds 10, while (S_1, P_1) allows the determination of the probability that the first toss results in 6, but (S_2, P_2) does not.

1.3 CONDITIONAL PROBABILITY AND INDEPENDENCE

Conditional Probability

Suppose that two six-sided fair dice are tossed and you learn that at least one of the two dice had resulted in 6. What is the probability now that the total of the numbers on the two dice is at least 10? Obviously, a revised probability should be larger than it was before you learned of the 6.

To answer this, consider the sample space $S = D \times D$, where $D = \{1, 2, \ldots, 6\}$, with the assignment of equal probabilities $1/36$ to each outcome (see Table 1.3.1). The event $A =$ [at least one 6] has 11 outcomes, and since you know that A has occurred, can serve as a new sample space, again with equal probabilities. However, these probabilities must be $1/11$ rather than $1/36$, in order to sum to 1. Let us refer

TABLE 1.3.1 Sample Space for Throw of Two Dice

	Second Die					
First Die	1	2	3	4	5	6
1						*a*
2						*a*
3						*a*
4						*ab*
5					*b*	*ab*
6	*a*	*a*	*a*	*ab*	*ab*	*ab*

to the new model as the *conditional model*, given A, and write the new probability of an event B as $P(B \mid A)$. For $B =$ [total of 10 or more] $= \{(4,6), (5,5), (5,6), (6,4), (6,5), (6,6)\}$, we get $P(B \mid A) = 5/11 = (5/36)/(11/36) = P(A \cap B)/P(A)$. This is larger than $P(B) = 6/36 = 1/6$.

Let a and b denote outcomes in A and B, respectively. Consider Example 1.2.1 with sample space $S = \{a, b, c, d, e, f\}$, with probabilities 0.30, 0.20, 0.25, 0.10, 0.10, 0.05. As before, let $A = \{a, b, d\}$ and $B = \{b, d, e\}$. If A is known to have occurred, then since $P(A) = 0.60$, we can form a new probability model with sample space A and revised probabilities $0.30/0.60 = 1/2$, $0.20/0.60 = 1/3$, and $0.10/0.60 = 1/6$. Since $A \cap B = \{b, d\}$, we find $P(B \mid A) = 1/3 + 1/6 = 1/2 = P(A \cap B)/P(A)$. Since $P(B) = 0.40$, the occurrence of A has again increased the probability of B.

We can avoid the need to define a new probability model by simply defining $P(B \mid A)$ for events A, B as follows.

Definition 1.3.1 For a given probability model (S, P), let A and B be two events with $P(A) > 0$. The *conditional probability* of B, given A, is $P(B \mid A) = P(A \cap B)/P(A)$. □

Consider Example 1.2.4. Since $P(R_1) = 0.6$, $P(R_1 \cap R_2) = 0.4$, we find that $P(R_2 \mid R_1) = 2/3$, while $P(R_2) = 0.5$. Rain on the first day makes it more likely that it will rain the second day. Similarly, $P(R_1 \mid R_2) = P(R_1 \cap R_2)/P(R_2) = 4/5 > P(R_1)$.

The definition $P(B \mid A) = P(A \cap B)/P(A)$ is useful in the form $P(A \cap B) = P(B \mid A)P(A)$, since in many cases conditional probabilities can be determined more easily from the fundamentals of the experiment than can the probabilities of intersections. In this form, conditional probabilities can be used to build probability models.

Example 1.3.1 Suppose that a sexual disease is present in 0.6% of 18- to 24-year-old men in a large city. A blood test for the disease is good, but not perfect, in the following way. The probability that a man with the disease is positive on the test is 0.98 (the *sensitivity* of the test). The probability that a man who does not have the disease is positive for the test is 0.01. (The *specificity* of the test is therefore 0.99.) What are:

(a) The probability that a man of that age selected randomly will be positive for the test?

(b) Given that such a man is positive for the test, what is the probability that he actually has the disease? The answer to this question may surprise you.

Let $S = \{n, d\} \times \{+, -\} = \{(d, +), (d, -), (n, +), (n, -)\}$, where d means that the man has the disease, n means that he does not, $+$ indicates that the test is positive, and $-$ indicates that the test is negative. Let $D =$ [man has disease] $= \{(d, +), (d, -)\}$, and $V =$ [test is positive] $= \{(d, +), (n, +)\}$. We are given

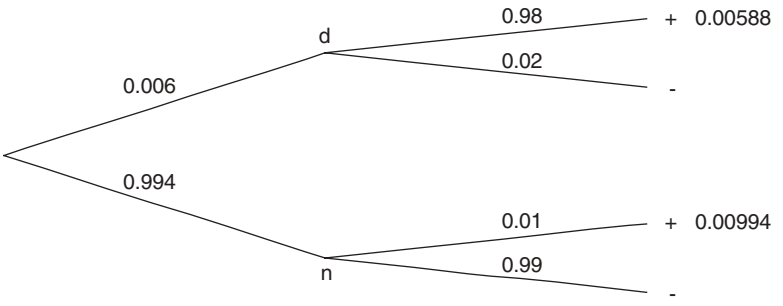

FIGURE 1.3.1 Tree diagram for sexual disease.

$P(D) = 0.006$, $P(V \mid D) = 0.98$, $P(V \mid D^c) = 0.01$. We want to determine $P(V)$ and $P(D \mid V)$.

Figure 1.3.1 presents these and other relevant unconditional and conditional probabilities. For example, $P((d, +)) = P(D \cap V) = P(D)P(V \mid D) = (0.006)(0.98) = 0.00588$. Similarly, $P((n, +)) = P(D^c \cap V) = P(D^c)P(V \mid D^c) = (0.994)(0.01) = 0.00994$. Since $V = (D \cap V) \cup (D^c \cap V)$ and the two events within parentheses are mutually exclusive, $P(V) = P(D \cap V) + P(D^c \cap V) = 0.00588 + 0.00994 = 0.01582$. Then $P(D|V) = P(D \cap V)/P(V) = 0.00588/0.01582 = 0.3717$. Only 37% of all those testing positive actually have the disease! This certainly suggests the retesting of those whose first test is positive. □

Example 1.3.2 A box contains r red and w white balls. Let $N = r + w$. Two balls are drawn consecutively and randomly without replacement. What is the probability that the second ball drawn is red?

We use two different approaches to answer the question. With luck the answers will be the same. Since the question concerns the ball chosen second, we choose a sample space in which the outcomes indicate the order in which the balls are chosen. Let R be the set of red balls and let W be the set of white balls. Let $B = R \cup W$. Let $S = \{(b_1, b_2) \mid b_1 \in B, b_2 \in B, b_1 \neq b_2\}$. Assign equal probability $1/N(S) = 1/[N(N-1)]$ to each outcome. Let $R_2 =$ [red on the second ball chosen]. Let us determine $N(R_2)$. For each possible choice of a red ball for the second ball chosen, there are $(N-1)$ choices for the first ball. Therefore, $N(R_2) = r(N-1)$ and $P(R_2) = [r(N-1)]/N(N-1) = r/N$, the proportion of red balls in the box. This is, of course, also the probability of the event R_1 that the first ball chosen is red.

Now consider the problem using conditional probability. Then $P(R_2) = P(R_1R_2) + P(R_1^cR_2) = P(R_1)P(R_2 \mid R_1) + P(R_1^c)P(R_2 \mid R_1^c) = (r/N)[(r-1)/(N-1)] + (w/N)[r/(N-1)] = [r/N(N-1)][(r-1) + w] = r/N = P(R_1)$. □

The problem many students have when first confronted with this type of example is caused by their difficulty in distinguishing between conditional and unconditional probabilities.

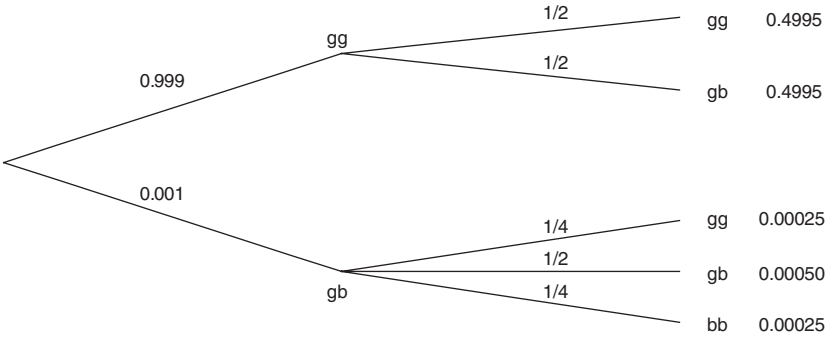

FIGURE 1.3.2 Tree diagram for gene analysis.

Let $A_1, A_2, \ldots, A_n$ be events. Then, applying the definition of conditional probability and canceling factors in the numerator and denominator, we get

$$P(A_1)P(A_2 \,|A_1)P(A_3 \mid A_1A_2)\cdots P(A_n \mid A_1A_2\cdots A_{n-1}) = P(A_1A_2\cdots A_n),$$

assuming that these conditional probabilities are all defined. For example, in consecutive random-without-replacement draws of four balls from a box with six red and seven white balls, the probability of the order (red–white–red–white) is $(6/13)(7/12)(5/11)(6/10)$.

Example 1.3.3 Suppose that a gene has two forms, g for good and b for bad. Every person carries a pair of genes, so there are three possible genotypes, gg, gb, and bb. Persons with genotype bb suffer from the disease and always die young, so they never have children. Persons of genotype gb do not suffer from the disease but are carriers in the sense that their offspring (children) may acquire the bad gene b from them. Suppose now that the proportions of people in the population of adults of genotypes gg and gb are 0.999 and 0.001. A female of known genotype gb has a child with a male drawn randomly from the adult male population (see Figure 1.3.2).

(a) What are the probabilities that a child of such a mating is of each of genotypes gg, gb, bb?
(b) Given that the child is of genotype gb, what is the probability that the male is of genotype gb?

If the male is of genotype gg, the child is equally likely to be gg or gb. If the male is of genotype gb, the child has probabilities $1/4$, $1/2$, $1/4$ of being gg, gb, or bb.

From Figure 1.3.2, the answers to part (a) are 0.49975, 0.50000, 0.00025. More generally, if p is the proportion of genotype gb in the population of adult males, the probabilities are $(1-p)/2 + p/4 = 1/2 - p/4$, $(1-p)/2 + p/2 = 1/2$, and $p/4$. In general, in answer to part (b), $P(\text{male is } gb \mid \text{offspring is } gb) = (p/2)/(1/2) = p$, so the conditional probability that the male is gb is the same as the probability that the child is gb. □

The examples for which we have used a tree diagram suggest useful identities. Suppose that a sample space can be partitioned into k disjoint subsets $A_1, \ldots, A_k$ whose probabilities, often called *prior probabilities,* are known and are positive. Suppose also that B is another event and that $P(B \mid A_i)$ is known for each i. Then:

Theorem 1.3.1 $P(B) = \sum_{i=1}^{k} P(A_i)P(B \mid A_i)$.

Proof: Since $B = BA_1 \cup \cdots \cup BA_k$, these events are disjoint, and $P(BA_i) = P(A_i)P(B \mid A_i)$, the identity follows from the additivity of P. □

The identity of Theorem 1.3.1 is sometimes called the *total probability formula*. From this formula and the definition of conditional probability, we get:

Theorem 1.3.2 (Bayes' Formula)

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(A_i)P(B \mid A_i)} \qquad \text{for } j = 1, \ldots, k.$$

The probabilities $P(A_j \mid B)$ are sometimes called *posterior probabilities*, since they are the revised probabilities [from the prior probabilities $P(A_j)$] of the events A_j given the occurrence of the event B.

Example 1.3.4 Boxes 1, 2, and 3 each have four balls, each ball being red or white. Box i has i red balls, $i = 1, 2, 3$. A six-sided fair die is tossed. If a 1 occurs, a ball is drawn from box 1. If a 2 or 3 occurs, a ball is drawn from box 2. If a 4, 5, or 6 occurs, a ball is drawn from box 3. What are (a) the probability that the ball drawn is red, and (b) the conditional probability that the ball was drawn from box j given that it was red?

Let A_i = [ball is drawn from box i] for $i = 1, 2, 3$. Let B = [ball drawn is red]. Then $P(A_j) = j/6$ and $P(B \mid A_j) = j/4$ for $j = 1, 2, 3$. Therefore, from Theorem 1.3.1, $P(B) = (1/6)(1/4) + (2/6)(2/4) + (3/6)(3/4) = 14/24$. From Bayes' theorem, $P(A_1 \mid B) = (1/24)/(14/24) = 1/14$, $P(A_2 \mid B) = (4/24)/(14/24) = 4/14$, and $P(A_3 \mid B) = (9/24)/(14/24) = 9/14$. The posterior probability $P(A_3 \mid B) = 9/14$ that the ball was drawn from box 3 given that it was red is larger than the prior probability $P(A_3) = 3/6$ that the ball would be drawn from box 3. □

Example 1.3.5 Your friend Zeke has been reasonably honest in the past, so that your prior evaluation of the probability that he is telling the truth when he claims to be tossing a fair coin rather than his two-headed coin is 0.9. The prior probability that he is tossing the two-headed coin is therefore 0.1. Zeke then tosses the coin n times and gets a head on every toss. What is the posterior probability that he tossed the fair coin?

Let F = [coin tossed is fair] and let B = [all tosses result in heads]. Then $P(B) = P(F)P(B|F) + P(F^c)P(B \mid F^c) = (0.9)(1/2^n) + (0.1)(1) = 0.9/2^n + 0.1$. Therefore, $P(F \mid B) = (0.9/2^n)/[0.9/2^n + 0.1] = 1/[1 + 2^n/9]$. As n becomes larger, the posterior probability that he is telling the truth goes rapidly to zero. Students can draw

their own conclusions about friends who tell them often of low-probability events that have occurred in their lives. □

Simpson's Paradox

In a paper appearing in the journal *Science*, Bickel et al. (1975) studied the rates of admission to graduate school by gender and department at the University of California at Berkeley. To make their point they invented the following data for the departments of "Machismatics" and "Social Warfare." For the combined departments their data were:

	Deny	Admit	Percentage
Men	300	250	45.5
Women	400	250	38.5

Assuming relatively equal ability among men and women, there seems to be discrimination against women. But the frequencies for the separate departments were:

	Machismatics			Social Warfare		
	Admit	Deny	Percentage	Admit	Deny	Percentage
Men	200	200	50.0	50	100	33.3
Women	100	100	50.0	150	300	33.3

Assigning equal probabilities $1/1200$ to each applicant and using obvious notation, with D_1 and D_2 denoting the events that the student applied to the Department of Machismatics and the Department of Social Warfare, we have $P(A \mid M) = 0.455$, $P(A \mid W) = 0.385$, $P(A \mid M \cap D_1) = 0.50$, $P(A \mid W \cap D_2) = 0.50$. When drawing conclusions about the relationships between variables, the tendency to "collapse" (combine) tables in this way leads to what is called *Simpson's paradox* (from a paper by E. H. Simpson, 1951, *not* named after the famous O. J. Simpson. In this case the department variable is called a *lurking variable*. Failure to take it into consideration leads to the wrong conclusion.

Independence

Consider the experiment in which a fair six-sided die is thrown twice. Let $D = \{1, \ldots, 6\}$, $S = D \times D$, and assign probability $1/36$ to each outcome in S. Let A be the event that the number appearing on the first throw is 5 or 6, and let B be the event that the number appearing on the second throw is at least 3. Then $P(A) = 12/36 = 1/3$, $P(B) = 24/36 = 2/3$, $P(AB) = 8/36 = 2/9$, and $P(B \mid A) = (2/9)/(1/3) = 2/3 = P(B)$. Thus, the occurrence of the event A does not affect the probability of the event B. This, of course, should seem intuitively clear, unless one believes that dice have memories.

We take this as a starting point in developing a definition of *independence* of events. Suppose that for two events A and B with $P(A) > 0$, $P(B \mid A) = P(B)$. Then

$$P(A \cap B) = P(B \mid A)P(A) = P(B)P(A),$$

and if $P(B) > 0$ so that if $P(A \mid B)$ is defined, $P(A \mid B) = P(A)$. Since

$$P(AB) = P(A)P(B) \tag{1.3.1}$$

implies both $P(B \mid A) = P(B)$ and $P(A \mid B) = P(A)$, and since (1.3.1) is symmetric in A and B and does not require either $P(A) > 0$ or $P(B) > 0$, we take (1.3.1) as the definition of the independence of two events.

Definition 1.3.2 Two events A and B are *independent* if $P(AB) = P(A)P(B)$. □

WARNING: Do not confuse the statement that two events A and B are independent with the statement that they are mutually exclusive (disjoint). In fact, if A and B are mutually exclusive, then $P(AB) = 0$, so that they cannot be independent unless at least one of them has probability zero.

It is easy to show that independence of A and B implies independence of the following pairs of events: (A, B^c), (A^c, B), (A^c, B^c). For example, $P(A^cB) = P(B) - P(AB) = P(B) - P(A)P(B) = P(B)[1 - P(A)] = P(B)P(A^c)$. In fact, independence is best thought of as a property of the probability measure on the sample space as applied to the partitioning of the sample space into four parts produced by the two events A and B. The fundamental idea of independence of events is used very often to build probability models.

Suppose that two experiments are to be performed, with corresponding probability models (S_1, P_1) and (S_2, P_2). Suppose also that it is reasonable to believe that the outcome of either experiment should not change the probability of any event in the other experiment. We can then produce a probability model for the combination of experiments as follows. Let $S = S_1 \times S_2$, and for $(s_1, s_2) \in S$, let $P((s_1, s_2)) = P_1(s_1)P_2(s_2)$. In this way we have defined a probability measure on S. To see this, note that for any events,

$$\begin{aligned} A_1 \subset S_1, A_2 \subset S_2, P(A_1 \times A_2) &= \sum_{s_1 \in A_1, s_2 \in A_2} P((s_1, s_2)) = \sum_{s_1 \in A_1} P_1(s_1) \sum_{s_2 \in A_2} P_2(s_2) \\ &= P_1(A_1)P_2(A_2). \end{aligned}$$

In particular, $P(S) = P(S_1 \times S_2) = P_1(S_1)P_2(S_2) = (1)(1) = 1$. Let $B_1 = A_1 \times S_2$ and $B_2 = S_1 \times A_2$, where $A_1 \subset S_1$ and $A_2 \subset S_2$. In the language of set theory, B_1 and B_2 are *cylinder sets.* B_1 is defined entirely in terms of the first experiment, B_2 entirely in terms of the second experiment. For the model (S, P), B_1

TABLE 1.3.2 Product Model

	e	f	g	sum
a	0.08	0.12	0.20	0.40
b	0.06	0.09	0.15	0.30
c	0.04	0.06	0.10	0.20
d	0.02	0.03	0.05	0.10
sum	0.20	0.30	0.50	

and B_2 are independent. Since $B_1 \cap B_2 = A_1 \times A_2$, $P(B_1 \cap B_2) = P(A_1 \times A_2) = P_1(A_1)P_2(A_2) = P(A_1 \times S_2)P(S_1 \times A_2) = P(B_1)P(B_2)$. (Pay attention to the subscripts or lack of subscripts on P!)

Example 1.3.6 Let $S_1 = \{a, b, c, d\}$ and let P_1 assign probabilities 0.4, 0.3, 0.2, and 0.1 to its outcomes. Let $S_2 = \{e, f, g\}$ and let P_2 assign probabilities 0.2, 0.3, and 0.5 to its outcomes. Then the outcomes of $S = S_1 \times S_2$ and the probability assignments under the independence model for the two experiments correspond to the 12 cells of Table 1.3.2.

Let $A_1 = \{b, c\}$, $A_2 = \{e, g\}$. Then the event $A_1 \times S_2$, the set of outcomes in the rows headed by b and c, has probability 0.50. The event $S_1 \times A_2$, the set of outcomes in the columns headed by e and f, has probability 0.70. The event $A_1 \times A_2$, the set of outcomes in the rectangle formed from the rows headed by b and c and the columns headed by e and f, has probability 0.35, which is, of course, $P(A_1 \times S_2)P(S_1 \times A_2) = (0.50)(0.70)$. □

In generalizing the property of independence for two events to that of three or more events $A_1, \ldots, A_n$, we want to be able to use the multiplication rule,

$$P(A_1^* \cap \cdots \cap A_n^*) = P(A_1^*) \cdots P(A_n^*), \tag{1.3.2}$$

where each A_i^* is either A_i or A_i^c. To assure this, it is not enough to require that these events be independent in pairs. It is equivalent that for any integer k, $1 \le k \le n$, and indices $1 \le i_1 < \cdots < i_k \le n$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}). \tag{1.3.3}$$

For example, for $n = 3$, we need $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$, $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$, and $P(A_2 \cap A_3) = P(A_2)P(A_3)$.

An inductive proof of the equivalence of (1.3.2) and (1.3.3) can be constructed, but we will avoid the messy details and instead show that (1.3.3) implies (1.3.2) for the special case of A_1, A_2, A_3^c. Since $(A_1 \cap A_2 \cap A_3^c) \cup (A_1 \cap A_2 \cap A_3) = A_1 \cap A_2$, $P(A_1 \cap A_2 \cap A_3^c) = P(A_1 \cap A_2) - P(A_1 \cap A_2 \cap A_3)$. From (1.3.3) this is $P(A_1)P(A_2)[1 - P(A_3)] = P(A_1)P(A_2)P(A_3^c)$.

Definition 1.3.3 Events $A_1, \ldots, A_n$ are *independent* if for any integer k, $1 \leq k \leq n$, and indices $1 \leq i_1 < \cdots < i_k \leq n$, (1.3.3) holds. □

Given two models (S_1, P_1) and (S_2, P_2) for two experiments and independence of these experiments, we constructed a product model (S, P), where $S = S_1 \times S_2$ and $P(A_1 \times A_2)$ for $A_1 \subset S_1, A_2 \subset S_2$. We can generalize this to the case of n models $(S_1, P_1), \ldots, (S_n, P_n)$. We let $S = S_1 \times \cdots \times S_n$, and for any outcomes $s_1 \in S_1, \ldots, s_n \in S_n$, assign probability $P_1(s_1) \ldots P_n(s_n)$ to $(s_1, \cdots, s_n) \in S$. Again we find that events which are defined in terms of nonoverlapping indices are independent. For example, for $n = 4$, independence of the events $A_1, \ldots, A_4$ implies the independence of $A_1 \cup A_3^c$, A_2, and A_4^c.

Example 1.3.7 Three new car salespeople, Abe, Betty, and Carl, are to be assigned to the next three customers. The three have differing sales skills: Abe makes a sale with probability 0.3, Betty with probability 0.2, and Carl with probability 0.1. If the three salespeople do or do not make sales independently, what is the probability that the three make a total of at least one sale?

Let A be the event that Abe makes a sale, and define B and C similarly for Betty and Carl. Then, since $P(\text{at least one sale}) = P(A \cup B \cup C) = 1 - P((A \cup B \cup C)^c) = 1 - P(A^c B^c C^c) = 1 - P(A^c)P(B^c)P(C^c) = 1 - (0.7)(0.8)(0.9) = 1 - 0.504 = 0.496$. □

Example 1.3.8 During the seventeenth century the French nobleman Antoine Gombauld, the Chevalier de Mere, a gambler, wrote to the mathematician Blaise Pascal concerning his experience throwing dice. He had been able to win regularly by betting that at least one 6 would appear in four rolls of a die. On the other hand, he was losing money when he bet that at least one double-6 would occur in 24 throws of two dice. It seemed to de Mere that he should have about the same chance of winning on each of these two bets.

It seemed "obvious" that the probability of at least one 6 should be $2/6$ for two throws of a die, $3/6$ for three throws, and so on. This reasoning seems to go bad for seven throws, however, so perhaps we need to think a bit more carefully. Using independence and DeMorgan's law, similarly to Example 1.3.5, for n throws of one die we get $P(\text{at least one 6}) = 1 - P(\text{no 6's}) = 1 - (5/6)^4 = 1 - 0.4823 = 0.5177 > 0.5$, so de Mere should have been a winner, although he had to be patient. On the other hand, for $n = 24$ throws of two dice, $P(\text{at least one 6–6}) = 1 - (35/36)^{24} = 1 - 0.5086 = 0.4914 < 0.5$, so that de Mere should have expected to lose, although slowly. To determine the difference in success rates between the two games experimentally, de Mere must have played very often and must have kept very good records. We have reason to be skeptical about de Mere's story. □

Example 1.3.9 Consider a system consisting of three components, 1, 2, and 3. Current (in the case that this is a wiring diagram with resistors 1, 2, and 3) or traffic (in the case that this is a system of highways with bridges 1, 2, and 3) must travel

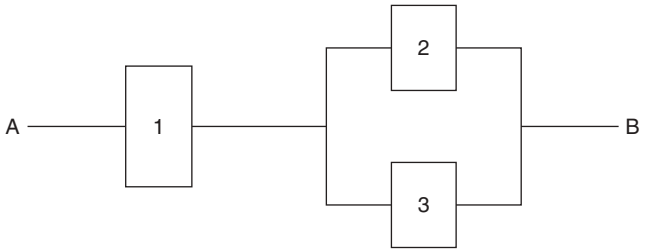

FIGURE 1.3.3 Reliability diagram.

from point A to point B. The system works if component 1 works and either of 2 or 3 works. Suppose that the three components have reliabilities (probabilities of working) 0.9, 0.8, and 0.7. Suppose also that the events that the three components function as they should are independent. What is the reliability of the system? That is, what is the probability that the system works? (See Figure 1.3.3.)

Let W_i for $i = 1, 2, 3$ be the probability that component i works. Then the reliability of the system $= P(W_1 \cap (W_2 \cup W_3)) = P(W_1)[1 - P(W_2^c)P(W_3^c)] = (0.9)[1 - (0.2)(0.3)] = 0.846$. □

Example 1.3.10 Our football team will play games against teams 1, 2, and 3. It has probabilities 0.7, 0.6, and 0.4 of winning each game, and it is reasonable to believe that the events of winning each game are independent. Given that the team wins at least one of the first two games, what is the conditional probability that it wins at least one of the last two?

Let $W_i =$ [our team wins the game with team i] for $i = 1, 2, 3$. We need to find $P(W_2 \cup W_3 \,|W_1 \cup W_2)$. A Venn diagram is helpful. Note that $(W_1 \cup W_2)(W_2 \cup W_3) = W_2 \cup (W_1 W_2^c W_3)$, so that using the independence of W_1, W_2, W_3, we get $P((W_1 \cup W_2)(W_2 \cup W_3)) = 0.6 + (0.7)(0.4)(0.4) = 0.712$. Since $P(W_1 \cup W_2) = 0.7 + 0.6 - (0.7)(0.6) = 0.88$, we get $P(W_2 \cup W_3 \,|W_1 \cup W_2) = 0.712/0.880 = 89/110$. What is the conditional probability that the team wins at least two games given that it wins at least one of the first two? □

Problems for Section 1.3

1.3.1 Let $S = \{a, b, c, d, e\}$ and let P assign probabilities 0.2, 0.3, 0.1, 0.3, and 0.1, respectively. Let $A = \{a, b, c\}$ and $B = \{b, c, d\}$. Find $P(A)$, $P(B)$, $P(A \mid B)$, $P(B \,|A)$, and $P(A^c \mid B^c)$.

1.3.2 A fair coin is tossed three times. Let $A =$ [at least one of the first two tosses is a head], $B =$ [same result on tosses 1 and 3], $C =$ [no heads], $D =$ [same result on tosses 1 and 2]. Among these four events there are six pairs. Which of these pairs are independent? Which are mutually exclusive?

1.3.3 A bowl contains five balls numbered 1, 2, 3, 4, 5. One ball is drawn randomly, that ball is replaced, balls with larger numbers are withdrawn, then a second ball is drawn randomly.