Elementary Probability

EDWARD O. THORP

Professor

Department of Mathematics

University of California, Irvine, California

ROBERT E. KRIEGER PUBLISHING COMPANY
HUNTINGTON, NEW YORK
1977

Original Edition 1966
Reprint 1977, with corrections

Printed and Published by
ROBERT E. KRIEGER PUBLISHING CO., INC.
645 NEW YORK AVENUE
HUNTINGTON, NEW YORK 11743

Printed in the United States of America

Library of Congress Cataloging in Publication Data

Thorp, Edward O.
Elementary probability.

Reprint of the edition published by J. Wiley, New York.
Bibliography: p.
Includes index.
1. Probabilities. I. Title.
[QA273.T44 1976] 519.2 75-45419
ISBN 0-88275-389-4

Preface

This book is designed to provide material for an introductory probability course for one quarter, one semester, or even two quarters. As a brief text in its own right, it is also intended to be integrated with the standard two years calculus course. Chapters 1 and 2 on finite probability require no calculus. Chapter 3 on discrete probability uses facts about absolutely convergent series, the geometric and e^x power series, and has some epsilon-delta proofs at the end. Chapter 4 on continuous probability uses integration techniques and the definition of continuous function. No knowledge of differentiation is required.

It divides naturally into these three major sections. Successive sections assume somewhat greater sophistication on the part of the student. The goal of the book is to provide core curriculum material in probability, and at the same time reinforce the calculus course. The diverse needs of students of probability suggest that after this core material is covered, the various major groups of users—social scientists, physical scientists, mathematicians, etc.—may wish to be given their own continuation course in probability and statistics, complete with applications.

This book can be used as the mathematical core of an introductory course in probability for those with special applications in mind, such as economists, psychologists, and biologists. In this case additional problems giving applications of the material to these areas may be desired. This is best left to be provided by those who work in the given area. The book *Mathematics for Psychologists—Examples and Problems* by Bush, Ableson, and Hyman, Scientific Research Council, New York, 1956, does this for psychologists.

There are many lengthier textbooks on probability at a variety of levels. This book, however, is not intended to compete with them nor is it intended to cover completely the "standard" introductory material.

The emphasis is on the conceptual development of the subject and in making each successive step seem to be the natural consequence of the preceding material. Thus, discrete probability is presented as a natural outgrowth of finite probability. Continuous probability is suggested by facets of the discrete theory.

The treatment of continuous probability is comparatively brief. I believe that the student of continuous probability should first acquire a higher level of competence in calculus than generally results from the introductory courses.

My treatment can be augmented quite simply in many ways. Some suggestions are:

1. A derivation of Stirling's formula and some applications. The treatment by H. Robbins, "A Remark On Stirling's Formula," *American Mathematical Monthly*, Vol. 62, 1955, pp. 26–29, is suggested.
2. A detailed elementary derivation of the normal approximation to the binomial distribution. The treatment in J. Neyman, *First Course in Probability and Statistics*, Holt, Rinehart and Winston, 1950, New York, pp. 234–242, is suggested and should follow Stirling's formula.
3. A proof of Weierstrass' theorem that a continuous function on a closed interval is the uniform limit of polynomials. This can be proven by probability techniques, using Bernstein polynomials. The proof is perhaps considerably easier than the usual one. A detailed treatment may be found in Robert G. Kuller, "Coin Tossing, Probability, and the Weierstrass Approximation Theorem," *Mathematics Magazine*, Vol. 37, No. 4, September 1964, pp. 262–265. The only fact which appears to be used, beyond elementary calculus and our probability treatment, is that a continuous function on an interval $a \leq t \leq b$ is bounded, which, of course, is generally assumed in elementary calculus courses.
4. Moment generating functions.
5. Graphs and trees.
6. Markov chains.

These topics and many others which were omitted were thought better left either to the discretion of the instructor or to a more specialized continuation course. (This book could be used for the first semester of a general introductory probability and statistics course. The second semester could consist of continuation courses for more specialized groups.)

I try to use modern terminology and notation, in conformity with present and future trends in undergraduate mathematics education.

Computer problems have been included to be used or not at the discretion of the instructor. The student who knows little or nothing about computers can organize these problems for machine solution, that is, write them up so that they are readily programmed. The student who has

had an introductory course in, say, Fortran can also complete the programming and obtain a machine solution. Programming courses will soon be standard for scientists and engineers.

The problems are divided into unstarred problems, starred problems, and problems or parts of problems involving machine computation. The unstarred problems are an essential part of understanding the material. The starred problems and the computation material are optional. The more difficult problems have been marked with a double star.

Terms that are defined informally are in boldface. Cross references to theorems, exercises, and so forth are given in "minimal" form. For instance, to refer to a theorem in Chapter 3, Section 1, we indicate Theorem 3 when within the section, Theorem 1.3 when elsewhere in the same chapter, and Theorem 3.1.3 when occurring in a different chapter.

Exercises and Suggested Reading generally appear at the end of the appropriate section, rather than at the end of the chapter. In addition to formal exercises, numerous verifications are suggested informally in the text. References in the text are only by author and title. The detailed references appear at the end of the book. It is helpful to place these references on reserve in the library for student use.

The suggested reading is of several kinds. There are selections from other books covering the same ground which can provide supplementary work for readers. Additional examples and exercises will be found in these selections. Furthermore, the reader should have a facility and familiarity with using and referring to the standard texts as part of his knowledge of probability. The level of difficulty of the supplementary material varies somewhat to satisfy varying levels of ability.

Historical, philosophical, and discursive selections are included in the suggested reading to indicate to the student how probability theory has developed and unfolded through the centuries, and may help to develop a unified view of the subject for the reader.

I wish to thank Professor Bernard Gelbaum for several helpful suggestions, and C. Joseph Frank and the other editors of John Wiley and Sons for their conscientious and thorough assistance. Corrections, comments, and suggestions from readers are appreciated.

EDWARD O. THORP

December 1965

Contents

1 Elementary Probability Notions 1

1. What is probability? 1
2. Equally likely alternatives 4
3. Elementary counting and combinatorial analysis 7
4. Appendix: Finite summation and product notation; mathematical induction 14

2 Probability Measures on Finite Sample Spaces 19

1. Elementary set theory 19
2. Functions 27
3. Probability measures on finite sample spaces 30
4. Probability measures and relative frequencies 33
5. Conditional probability and independence 36
6. Bayes' rule 40
7. Independent trials 43
8. The probability of the union of n events 47

3 Random Variables on Discrete Sample Spaces 50

1. Discrete sample spaces 50
2. Random variables 55
3. The expectation of a random variable 66

4. Moments, variance, and standard deviation 73
5. Chebychev's inequality and "the law of averages" 79
6. Cartesian products of infinitely many discrete spaces 85
7. The Poisson distribution and applications 89

4 Continuous Probability 97

1. Introduction 97
2. Borel sets on the line and continuous density functions 103
3. The normal approximation to the binomial probability law 111
4. Probability and measure theory 123

REFERENCES 127
SOLUTIONS TO EXERCISES 129
INDEX 147

CHAPTER 1

Elementary Probability Notions

1. WHAT IS PROBABILITY?

A good dictionary will give several distinct meanings for the word probability. We use the term to mean the mathematical theory of probability. We begin with a brief nonmathematical description of how this theory originated and what it has become today.

Since earliest times, men have tried to explain their environment and to use their explanations to control their future. For example, some Southwestern Indian tribes believed that "spirits" caused rainfall. To bring rain after drought, they performed ceremonial rain dances for the benefit of these spirits.

Man's efforts to explain his environment finally produced the very successful scientific method. One of the early spectacular triumphs of this method was Sir Isaac Newton's* law of universal gravitation. The law of gravitation and its consequences were worked out during the succeeding centuries, giving rise to the branch of physics known as Newtonian mechanics. For example, the past, present, and future positions of the principal bodies of the solar system were determined for thousands of years with great accuracy.

The results of Newtonian mechanics were so satisfactory and extensive that scientists began to feel that the universe was one vast machine whose past, present, and future could be completely predicted if enough data were supplied. The great French scientist and mathematician, Pierre

* In his early twenties, Newton (1642–1715) scored two of the greatest scientific achievements of all time. He invented the calculus and he formulated the law of universal gravitation. (Leibnitz later discovered the calculus independently. It is his notation that was used until recently.) Newton made many other contributions to mathematics and to physics, among them the binomial theorem and the discovery that a prism separates sunlight into many colors. Newton is ranked as one of the three greatest mathematicians of history by E. T. Bell in his readable and informative *Men of Mathematics*. An excellent biography is *Sir Isaac Newton* by Louis Trenchard More.

Simon Laplace, said that if the positions and velocities of all the particles in the universe for some one instant of time were given to an intelligence great enough to perform the required computations (Laplace's famous divine calculator), the past, present, and future of the universe for all time could be determined.

This view of the universe as one vast predetermined machine dominated scientific thought until this century. It gave rise to the philosophy of determinism (and the endless debates about "determinism" versus "free will").

Despite the triumphs of Newtonian mechanics, it did not give practical answers in most situations. Suppose, for example, that we flip a coin and let it come to rest on a horizontal surface. Which side will be uppermost? Newtonian mechanics argues that we could calculate the answer to our question if we had enough data about the state of affairs when the coin was launched. But it is not feasible to get sufficient data. Thus in practice we regard the answer to our question as somewhat uncertain. Attempts to analyze such "experiments" with uncertain outcome by the use of mathematics led to the development of probability theory.

For purposes of analysis one might regard all phenomena as being composed of two parts: a predictable Newtonian deterministic part and a chaotic, uncertain chance or probabilistic part. Even in the most predictable of phenomena some of this chance part is always present, in the form of experimental error.

A natural argument for a Newtonian is that the chance part of phenomena will continue to become predictable as science and knowledge advance. At the end of the last century the view was common that future physics would simply consist of determining quantities to successively greater numbers of decimal places.

The discovery of the radioactive decay of atoms demolished this view with startling suddenness. Certain types of atoms were found to "explode" and the times of the explosions were unpredictable. The old theories did not explain it. The new quantum theory of matter was developed to do so. The quantum theory turns out to be intrinsically probabilistic, that is, insofar as the quantum theory goes there is no certainty or determinism in the universe. The future cannot be calculated, no matter how much data we have about the past and the present. Only the probabilities of various possible futures can, in principle, be determined. The apparent determinism that we see on the macroscopic (roughly, everyday world) level is an illusion. Deterministic laws work well there simply because of certain laws of the theory of probability, such as the law of large numbers. We prove a simple form of this theorem later in the text.

Today, man's view of the physical universe is fundamentally probabilistic, not deterministic.*

The first known attempts to analyze chance phenomena mathematically were by Gerolamo Cardano (1501–1576).† At the time, Cardano was a graduate student in mathematics at the Italian University of Padua. He was also an active gambler. He computed the odds when various numbers and kinds of dice are rolled. Cardano wrote up his results in what might be called the first practical handbook for gamblers, but the manuscript lay unpublished until 1663. Thus it happened that two mathematical eminents, Blaise Pascal‡ and Pierre Fermat,§ gave the theory of probability its initial impetus. They began a correspondence in 1654 concerning dice questions raised by a gambler, the Chevalier de Méré. The techniques and methods developed by them initiated the theory.

Probability theory today has spread far beyond applications to games of chance. A catalogue of most of the principal developments through 1865 is given by Todhunter. Though this account is voluminous, most developments in the theory have occurred recently. Probability and its applications, primarily statistics, are among the major areas of activity in modern mathematics today.‖

Genetics, astronomy, industrial quality control, reliability studies of space vehicles, the theory of experimental errors, educational testing and psychological measurements, public opinion polls and market research, agricultural experiments and crop testing, and business and economic

* A minority of physicists, notably the late Albert Einstein (1879–1955), believe that there are more detailed deterministic laws at the subatomic, or even subelectronic, level which will be found to explain the quantum phenomena.

† Gerolamo Cardano, *Book on Games of Chance* (written about 1520). For a translation of the book by Sidney H. Gould and an extensive discussion, analysis and comments, see *Cardano, The Gambling Scholar*, by Øystein Ore.

‡ (1623–1662) Pascal contributed to probability, projective geometry, and invented the first "desk calculator." He is well known to students of philosophy and literature for his prose works. Pascal also contributed to physics.

§ Fermat (1601–1665) was a lawyer and an amateur mathematician. He made several notable contributions to mathematics. The most famous and enigmatic is "Fermat's last theorem" which asserts that the equation $x^n + y^n = z^n$ has no nontrivial solution in positive integers when n is greater than 2. Fermat wrote in the margin of a book that he had discovered a "truly marvelous" proof of this theorem which (alas!) the margin was too small to contain. For three centuries, mathematicians have sought to prove or disprove this theorem. It has been found to be true whenever n or each of the numbers x, y, and z, is less than huge. No one yet knows whether the entire theorem is true or not.

‖ Mathematicians are often themselves surprised to learn that roughly one-eighth of today's mathematical activity, as measured by space used in recent issues of *Mathematical Reviews*, is in probability and statistics.

predicting are only a few of the diverse fields in which probability and its applications are important.

Today knowledge of the elementary theory of probability is necessary for engineers and for workers, not only in the physical sciences but also in the social and biological sciences. Students of business and economics, and even the well-informed business man, also need an introduction to probability ideas. Our discussion of determinism versus chance shows that the philosopher needs to understand probability. Finally, some grasp of probability is essential for the well-informed layman who wants to appreciate the ubiquitous impact of probability and statistics in our rapidly changing world.

SUGGESTED READINGS

Feller, Introduction.
Newman, comments on Cardano on page 119 and page 1395 fn.

2. EQUALLY LIKELY ALTERNATIVES

In probability theory we shall be concerned with uncertainty, specifically with situations which have more than one possible outcome. For example, when a die is rolled, any one of six faces may finally be uppermost; there are six possible outcomes. This situation is called an *experiment*. The outcomes of an experiment are called *events*.

Some events cannot be thought of as a list of two or more other events. In the rolling of a die, the event that a specified face comes up is such an event. These simplest or "atomic" (not further decomposable, as the fathers of chemistry conceived of atoms) events are called elementary events. Events which can be thought of as made up of, or decomposable into, simpler events are called compound events. In rolling a die, an example of a compound event is the event "an even (numbered) face comes up." Another example is the event "face one or face two comes up."

We wish to discuss the various likelihoods or probabilities of events in some precise way. In particular we wish to assign numbers to events in some "natural" way. We begin (as did Laplace) with experiments having the following particularly simple characteristics.

(*a*) There are only a finite number of possible elementary outcomes.
(*b*) The elementary outcomes are assumed to be equally likely.

Example 1. If a coin is flipped and allowed to come to rest on a horizontal surface, there are two elementary outcomes, "heads" and "tails." (We neglect the remote possibility that the coin rolls away or stands on edge. We could agree to toss the coin over again if and when this occurs.) The coin is called "fair" or "true" if we assume that heads and tails are equally likely. Otherwise it is called "biased."

Example 2. A die is rolled and allowed to come to rest on a horizontal surface. There are six elementary outcomes. The die is called "fair" or "true" if the six outcomes are assumed to be equally likely. Otherwise it is biased. In earlier times, tetrahedral dice (having four congruent equilateral triangles as faces) were common. They are called astragals. (Note that there are five regular geometric solids, all suitable for use as dice.)

Example 3. A pack of N cards, labeled from 1 to N is shuffled and a card is then drawn. There are N elementary outcomes. We generally assume that the outcomes are equally likely. If we say that the shuffling is "thorough" or "random" or that the pack is "well-shuffled," then in particular we are assuming that these outcomes are equally likely. But random shuffling means still more. It means that all possible final arrangements or permutations are equally likely. We discuss this in Section 3.

It is clear from the examples that we may assume that for any positive integer N there is an experiment with N equally likely elementary outcomes.

In each of the preceding examples, we say "if we *assume* that the outcomes are equally likely." In fact, all real coins and dice have defects and biases. It can be demonstrated theoretically that even card shuffling, which most people consider very good, is not random.

The situation is analogous to that in plane geometry. Recall that no real "points"—such as the dot you make on your paper or a tiny speck viewed with an electron microscope—have the qualities we imagine for geometric points. The points of plane geometry have a precise position but no length, width, or breadth. The "points" in the real world that they represent for us have dimensions (size) and consequently do not have a precise position. Similarly, the (straight) lines of geometry have one dimension, length, but no width or height. Again, the "lines" of the real world have all three. Even the solids of geometry have precise boundaries. But solid objects of the real world do not.

Plane geometry is the prototype of a mathematical system. It has axioms (or postulates) that are assumed as given without proof (such as "through every two points there is precisely one straight line"). These axioms are about certain undefined terms (such as "point" and "straight line"). New terms are defined by using the rules of logic and the undefined terms and any previously defined terms. For instance, triangle is defined by using "point" and "straight line." Theorems are proven by using the rules of logic, the axioms, and any previously proven theorems.

Corresponding to the points, lines, triangles, and other terms of plane geometry, most of us have a fairly extensive set of experiences of "real" points, lines, and triangles. This makes it easier for us to "visualize" and work with the concepts of plane geometry. Furthermore, the rules that real objects satisfy turn out to resemble closely the rules which the corresponding abstract terms of geometry satisfy. If, for example, we measure the angles in a real triangle, we find that they add to very nearly 180°. Discrepancies from theory can generally be explained as "experimental error" or as due to imperfections in the triangle.

Plane geometry yields conclusions which can be translated into statements about real world situations. For this reason, it is called a mathematical model for these situations. Since the statements about the real world which are deduced from geometry correspond well with what is actually observed, the model is "good." In the same way, the theory of probability has been developed as a mathematical system with axioms, undefined terms, defined terms, theorems, etc., as a model for uncertain or "chance" phenomena. One can proceed on a purely deductive basis, just as with plane geometry. The justification of the system, and also of particular assumptions such as the one that states certain events are equally likely, is that the results of the model correspond closely with those observed in reality.

Most people today have a fair geometric intuition which gives them insight into the formalism of geometry. But most people today have little if any probabilistic intuition.* Part of the job of an elementary course is to develop the reader's intuition and experience with "chance" situations. Therefore we shall give numerous examples and illustrations of real world situations which correspond to statements of the theory. Advanced courses in probability generally use a more formal approach.

Now we return to the problem of assigning numerical likelihoods or probabilities to events. Suppose an event E consists of m of the n equally likely elementary events of an experiment. Then we say, as did Laplace, that the probability $P(E)$ of the event E is the number m/n. For example, the probability that an even face will come up when a true die is rolled is 3/6 or 1/2.† Heads and tails each have probability $\frac{1}{2}$.

We have given a rule for assigning to each event E a number $P(E)$, the probability of the event E. Notice that for each event E we have $0 \leq P(E) \leq 1$. Now $P(E) = 1$ if and only if E is the event "any of the elementary outcomes." Can $P(E) = 0$? It seems reasonable to say $P(E) = 0$ precisely when E is an event which includes *none* of the elementary outcomes. Since we have assumed that one of the outcomes must occur, these E are

* There will no doubt come a time in social development when man's probabilistic intuition will be nearly as well developed as his geometric intuition.

† We all know that $3/6 = 1/2$. The point of this is to show you where the 1/2 came from.

precisely the impossible events. Why concern ourselves with impossible events? We shall see later that the concept of an impossible event is valuable in much the same way that "nothing" (later known as zero) was valuable in the evolution of our number system.

EXERCISES

2.1 In each of the following experiments, describe the elementary outcomes and discuss whether or not you think that they are equally likely.

(*a*) An urn is filled with n identical balls numbered from 1 to n. The balls are thoroughly mixed. A ball is selected "blindly" and the number is recorded.

(*b*) A person steps on a scale which tells weight to the nearest pound from 0 to 250 lb.

(*c*) A wheel of fortune is spun and one of 48 numbers on the wheel comes up.

(*d*) Five cards are selected from an ordinary well-shuffled deck and the number of Aces obtained is recorded.

(*e*) A coin is tossed until the same result occurs twice in succession.

(*f*) A bucket of n dice are simultaneously rolled and the total is recorded.

(*g*) A coin is tossed two times and the successive results are listed *in order*.

(*h*) A coin is tossed n times and the successive results are listed in order.

(*i*) A book is selected "at random" from a library shelf, opened "at random" to a page, and the number of misprints is recorded.

(*j*) An urn contains four balls numbered from 1 to 4. The balls are thoroughly mixed, two are selected "blindly," and their numbers are recorded.

2.2 A group of n men and n women can be regrouped into couples in how many distinct ways? (*Hint:* Begin by solving the problem for $n = 1$, $n = 2$, and $n = 3$.)

2.3 (*a*) How many 5 digit numbers are there?

(*b*) How many 5 digit numbers are there with all digits distinct?

(*c*) What fraction of the 5 digit numbers have all digits distinct?

Solve the above with each of the following interpretations of a 5 digit number: The number in standard form has precisely 5 digits (for example, 00003 is a one digit number) and the number has at most 5 digits (for example, 00003 is a 5 digit number).

Be able to solve the above for n digits and for various number base systems.

2.4 A certain typewriter has 48 keys. A monkey hits the keys completely at random twenty-nine times. Assuming the keys are distinct, how many different expressions are there that the monkey could have typed? What is the probability of each of these expressions?

3. ELEMENTARY COUNTING AND COMBINATORIAL ANALYSIS*

In actual games of dice two or more dice are generally rolled. The total of the numbers of the uppermost faces is of interest. In the game of

* Readers with comparatively weak backgrounds are advised to study the background material on summation and induction in the next section *before* reading this section.

"craps," for example, precisely two dice are rolled. The possible totals are readily seen to range from 2 to 12 inclusive. What are the probabilities of the various totals?

There are six ways in which the first die can come up and six ways in which the second die can come up. Thus there are 6×6, or 36 ways in which the two dice can come up. Notice that we have considered as distinct, pairs of events such as "the first die shows a three and the second die shows a five" and the reverse, "the first die shows a five and the second die shows a three." If each die is fair, it seems plausible to suppose that each of the 36 ways is equally likely. We make this assumption. (It is justified further when we discuss independence in Section 2.7.) A routine listing of these cases follows.

Value of total:	2,12	3,11	4,10	5,9	6,8	7	Sum
Ways each:	1	2	3	4	5	6	36
Probability:	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	1

The craps example suggests the following useful general principle.

Theorem 1. If k experiments are performed and there are n_i different outcomes for the ith experiment, where i ranges from 1 to k, then there are $n_1 \times \cdots \times n_k$ or $\prod_{i=1}^{k} n_i$ different outcomes when all k experiments are performed. (In the craps example, $k = 2$ and $n_1 = n_2 = 6$.)

Again it is understood that order is important. Suppose $E_1, \cdots, E_k$ are the actual outcomes of experiments $1, \cdots, k$, respectively. Let the k-tuple or list $(E_1, \cdots, E_k)$ represent the corresponding outcome when all k experiments are performed. Then two such outcomes $E = (E_1, \cdots, E_k)$ and $F = (F_1, \cdots, F_k)$ are identical if and only if $E_i = F_i$ for all i from 1 to k. It may happen that the F_i are the same as the E_i rearranged, that is, in a different order. For example, it might happen that $E_1 = F_k$, $E_2 = F_{k-1}$, and so forth. In this case the outcomes E and F are to be considered distinct.

To see that the theorem is true, we argue that the number of outcomes is the same as the number of (distinct, ordered) k-tuples. We get each such k-tuple once and only once by making any of n_i choices for E_i as i ranges from 1 to k. This yields $\prod_{i=1}^{k} n_i$ k-tuples.

One of the important procedures for the application of probability to practical problems involves sampling. Sampling refers to performing a series of experiments and listing the observed outcomes. The list is called a **sample**.

Example 1. *Ordered sampling with replacement.* Suppose an urn contains n balls numbered inclusively from 1 to n. Consider k experiments, each of which consists of selecting a ball, recording its number, and replacing it. The urn of balls could be replaced by any finite "population" of "individuals." Our procedure is called **sampling with replacement.** If we list the results of the experiments in order from 1 to k, then it is **ordered sampling** and the list is an **ordered sample.** Theorem 1 applies and tells us that there are n^k such distinct ordered samples.

Suppose instead that we did not replace the balls after they were drawn. Then the procedure is termed **ordered sampling without replacement.** Clearly we require $1 \leq k \leq n$. Then we have the following.

Theorem 2. For a population of n elements there are $n(n-1)\cdots(n-k+1)$ distinct ordered samples without replacement of size k. In particular the number of distinct orderings (arrangements, permutations) of n elements is $n(n-1)\cdots 2 \cdot 1$.

The argument is similar to the one for Theorem 1. There are n ways to fill the first "slot" of the k-tuple. Since we do not replace the element so used, there remain only $n - 1$ elements to choose from to fill the second slot. Continuing, we have the result. The expressions in the theorem recur frequently, so we introduce the abbreviations $n(n-1)\cdots 2 \cdot 1 = n!$ (read "n factorial") and $n(n-1)\cdots(n-k+1) = (n)_k$. For $k > n$, $(n)_k = 0$.

Example 2. Suppose that five true dice are rolled. What is the probability that no two show the same face? There are 6^5 equally likely outcomes and in $(6)_5$ of these, the faces of the dice are all different. Thus the probability is $(6)_5/6^5 = 5/54$.

Next we consider **unordered sampling without replacement** from a population of n elements. Here we agree that two samples are the same if and only if the two samples have the same members, regardless of arrangement or order.

Example 3. Suppose that an urn contains three balls and that we draw two balls. (*a*) First let us regard the two balls as an ordered sample without replacement. Then the six possible elementary outcomes are (1,2), (1,3), (2,1), (2,3), (3,1), and (3,2). If they are equally likely, then each has probability 1/6. The probability of drawing a specific ball, say ball two, is 2/3. (*b*) If instead the sample is considered to be unordered, then we consider (1,2) and (2,1) as identical outcomes. The same holds for (1,3) and (3,1) and also for (2,3) and (3,2). Thus with unordered sampling there are only three elementary outcomes, each with probability $2/6 = 1/3$. Note that the probability that a specific ball is drawn is still 2/3.

Now let us consider the general case of unordered sampling without replacement. For an ordered sample of size k, there were $(n)_k$ different outcomes. Each of these outcomes can be rearranged in precisely $k!$ ways to produce other outcomes which are distinct for ordered sampling and identical for unordered sampling. Thus, in unordered sampling without replacement the elementary outcomes are simply nonoverlapping sets of elementary outcomes from ordered sampling without replacement. Since each of these sets is of size $k!$, we have the following.

Theorem 3. There are $(n)_k/k!$ elementary outcomes in unordered sampling without replacement.

Note that if the elementary outcomes of ordered sampling without replacement are equally likely, then each of the elementary outcomes of unordered sampling without replacement have probability $k!/(n)_k$ and so are also equally likely. (The converse is false. See Exercise 2.4.4.)

It is also of interest to discuss sampling when the outcomes are no longer equally likely. To distinguish, we give the following definition.

Definition 1. To refer to sampling, with or without replacement (whether ordered or unordered), in which the n^k or $(n)_k$ outcomes of ordered sampling are assumed to be equally likely, we use the term random sampling. The sample is called a random sample.

The expression $(n)_k/k!$ of Theorem 3 is an extremely important one in many branches of mathematics. We denote it by $b(n,k)$ and also by the traditional $\binom{n}{k}$. Observe that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, when $0 < k < n$. The expression $b(n,k)$ is called a binomial coefficient, because of its early discovery in connection with the binomial theorem. This theorem states that $(x + y)^n = \sum_{k=0}^{n} b(n,k)x^{n-k}y^k$, where $b(n,0)$ and $b(n,n)$ are given the values 1. In order that $n!/0!n!$ should give the right answers in these cases, we are led to define $0! = 1$. Consistent with $(n)_k = 0$ for $k > n$, we have $b(n,k) = 0$ when $k > n$. A common expression for $b(n,k)$ is the number of combinations of n things taken k at a time.

Example 4. How many different hands are there in the card game of bridge? Assume that all hands are equally likely. What is the probability that a specified hand contains no spades? (The probability that some hand on the table contains no spades is of course considerably larger.)

Bridge is a game played by four people known as north (N), south (S), east (E), and west (W). North and south are partners, as are east and west. The game is played with an ordinary 52-card pack. To begin the game, the deck is shuffled and 13 cards are dealt to each player. Such a set of 13 cards is called a **hand**. The whole configuration of four groups of 13 cards is a deal.

The first part of the problem reduces to finding the number of ways in which 13 cards can be selected from 52. This is $b(52,13)$. Hands with no spades are obtained by selecting 13 cards from 39. There are $b(39,13)$ such hands. Thus the probability that a hand contains no spades is $b(39,13)/b(52,13)$ or approximately 1.28×10^{-2}.

Such computations arise frequently in probability theory. It is instructive to compare the computation by hand (there is much canceling), by a desk calculator, by a computing machine—an easy Fortran program will suffice), and by reference to a table of binomial coefficients. One of the simplest and easiest ways to do hand calculations of expressions involving binomial coefficients or factorials is with the aid of a table of logarithms of factorials. This and the other tables we use in this text are conveniently available in *Handbook of Mathematical Tables*, First Edition (or later), *Supplement to Handbook of Chemistry and Physics*, Chemical Rubber Publishing Company, Cleveland, Ohio (1962).

SUGGESTED READINGS

Kemeny, Schleifer, Snell, and Thompson, Sections 10 and 11 (an introduction to flow diagrams).

Feller, Chapter II, Sections 1–4, and selected exercises from Section 10 give further material on Section 3. For binomial coefficients, in Chapter II, read Section 8 and the appropriate exercises from Section 12.

Bell, Chapter 5 (Pascal) for the beginnings of probability theory. Chapter 6 (Newton) is worthwhile reading because every educated person should have a rudimentary knowledge of the life and attainments of one of mankind's few unsurpassed intellects.

EXERCISES

3.1 (*a*) How many lines are determined by $n \geq 3$ points, no three of which are collinear?

(*b*) How many planes are determined by $n \geq 4$ points, no four of which are coplanar?

3.2 (*a*) Suppose a certain state prints auto license plates from a collection of 34 symbols (the alphabet, except for O and I, and the ten digits). How many different such license plates are there consisting of six or less consecutive symbols?

(*b*) If only the ten digits are used, how many different such license plates are there consisting of six or less consecutive digits?

3.3 Assume that the initial digit of a phone number is not zero (which gets "operator"), and that all other possible phone numbers are permissible.

(*a*) How many seven digits numbers are there?

(*b*) How many ten digits numbers are there?

3.4 At a certain university dance, a total of 83 unescorted men and 68 unescorted women attend. Suppose we are concerned only with whether or not various men and women pair off during the evening. Assume that each person is involved in at most one such pairing, and that the maximum number of couples need not be formed.

(*a*) In how many ways can precisely k couples be formed, $0 \le k \le 68$?
(*b*) In how many ways can the evening end?
(*c*) Generalize to m men and w women.

3.5 What is the probability that the last eight cards in a well-shuffled-deck consists of the ranks A, K, Q, J, 10? (See E. O. Thorp, *Beat The Dealer*, Revised, page 119.)

3.6 Verify the following recursion relations for binomial coefficients:* $b(n,k) + b(n, k+1) = b(n+1, k+1)$, $n \ge 1$, $0 \le k < n$, and $b(n, k+1) = (n-k)b(n,k)/(k+1)$, $n \ge 1$, $0 \le k < n$.

Computation: Make a flow chart for the computation of a table of binomial coefficients. Such a table is known as Pascal's triangle. See Bell, pages 87 and 88, for background. Consider two kinds of tables, those which start with $n = 1$ and continue to some $n = N$ and those which start at $n = M \ge 1$ and continue to some $n = N > M$. Note that overflow problems soon arise, that is, the answers are longer numbers than those the machine ordinarily stores.

3.7 Show that for $n \ge 1$, $1 - b(n,1) + b(n,2) - \cdots = 0$ and that $1 + b(n,1) + b(n,2) + \cdots = 2^n$.

3.8 If a fair coin is tossed n times (or if n fair coins are tossed), and all ordered sequences of outcomes are equally likely, show that the probability of k heads is $b(n,k)/2^n$. Show that the probability of an even number of heads and the probability of an odd number of heads are each $\frac{1}{2}$. (*Hint:* Use the result of the preceding exercise.)

3.9 How many different deals are there at bridge? Two deals having the same four hands, but dealt to different players, are to be considered as distinct.

3.10 We have discussed ordered sampling both with and without replacement, and unordered sampling without replacement. Conspicuously absent is unordered sampling with replacement. This is because the calculations are considerably more complex in general and the theory is less succinct. Discuss unordered sampling with replacement for $n = 2$, $k \ge 1$. Discuss it also for $n = 3$, $k = 3$.

3.11 Suppose that an urn contains four balls and that we draw two of them. List the distinct outcomes for each of the following.

(*a*) Ordered sampling without replacement.
(*b*) Ordered sampling with replacement.
(*c*) Unordered sampling without replacement.
(*d*) Unordered sampling with replacement.

***3.12** *Five-card draw poker.* An ordinary 52-card pack of cards is well-shuffled, and five cards are dealt to each of two or more players. The five cards dealt to a player are a **hand**. Hands are ranked, with higher ranking hands

* The second equation involves an expression of the form ab/c. Ordinarily, multiplication and division have equal force. One does not take priority over the other. Therefore ab/c might mean either $(ab)/c$ or $a(b/c)$. These expressions happen to be equal so there is no ambiguity. But ab/cd could mean $(ab)/(cd)$ or $[(ab)/c]d$ and these are generally different. To simplify the printing of fractions, we adopt the convention so beloved in the printing profession. Multiplication has priority over division. Thus ab/cd means $(ab)/(cd)$. This allows us to print on one line many fractions that would otherwise occupy two lines.

winning over lower ranking hands. Cards rank downwards in order, Ace, King, Queen, Jack, Ten, . . . , Two, (Ace)—an Ace may also be counted as low in the formation of a straight. A **straight** is a hand in sequence, such as Jack, Ten, Nine, Eight, Seven. A **flush** is a hand of one suit. **One pair** is a hand in which precisely two cards are of equal rank, such as 7, 7, 8, J, Q. A **pair** has the form a, a, b, c, d. **Two pairs** is a hand of the form a, a, b, b, c. **Three of a kind** is a hand of the form a, a, a, b, c. **Four of a kind** has the form a, a, a, a, b. A **full house** has the form a, a, a, b, b. A straight **flush** is a straight of one suit. **Nothing** is a hand which is none of the preceding.

In comparing hands of different types, we have in order of decreasing rank: straight flush, four of a kind, full house, flush, straight, three of a kind, two pairs, one pair, nothing.

Compute the probability of being dealt each of these types of hands.

3.13 *Equivalence classes of poker hands.* To compare two poker hands of the same type, the hand with the highest ranking cards wins. In comparing full houses, the three of a kind hand is compared first. In comparing one pair hands, the pairs are compared first. In comparing two pair hands, the highest ranking pairs are compared first. If they are equal, the remaining pairs are compared. If they are also equal, the lone remaining card of each hand is compared. Ties are possible in each of the types of hands. How many distinct hands, that is, hands of different rank, are there in five-card draw poker?

The $b(52,5)$ possible hands can be grouped into subsets such that two hands in any one subset are equal, or tie each other. Then when two of these subsets are compared, all the hands in one of them beat all the hands in the other. We refer to these subsets as equivalence classes. The problem is to find out how many of these equivalence classes there are.

3.14 Which equivalence class or classes determines the median poker hand? There are two possibilities. The first possibility is that there is one equivalence class with the probability of less than $\frac{1}{2}$, that a hand is worse, and also the probability of less than $\frac{1}{2}$ that a hand is better. If this is the case, find this equivalence class. The second possibility is that there is no equivalence class with the above property, in which case the totality of all hands divides into two sets, each with probability $\frac{1}{2}$—all the hands in one of the two sets are worse than all the hands in the other. If this second possibility occurs, find the two sets.

3.15 *Bridge.* Find the following probabilities.

(*a*) North is dealt the Ace of Spades.
(*b*) North is dealt no Spade.
(*c*) North is dealt all Spades.
(*d*) North is dealt all 13 of one suit.

3.16 *Computer determination of the dealer's probabilities in casino blackjack.* The casino game of blackjack is played with one (or more) bridge deck(s). The cards are shuffled and two are dealt to each player. Two cards are dealt to the dealer, one face up and one face down. The players in turn, and then the dealer, attempt to reach a total as close to 21 as possible without exceeding it. An Ace counts 1 or 11, 2 to 9 count their face value, and 10, J, Q, K each count 10. The dealer must draw to a total of 16 or less and must *stand* (no further drawing) if he has a total of 17 or more. If he can count an Ace as 11 to get a total between 17 and 21, he must do so and stand.

For each possible dealer's up-card, use a computer to determine the probabilities of the various dealer's totals: 17, 18, 19, 20, 21, and over 21. The easiest case, a dealer's up-card of 10, can be checked readily by hand.

These probabilities are of central importance in the analysis of the game. Further details and the table of dealer's probabilities appear in Thorp, *Beat The Dealer*.

4. APPENDIX: FINITE SUMMATION AND PRODUCT NOTATION; MATHEMATICAL INDUCTION

In the last section we used a special abbreviation for the finite sum of terms which appears in the binomial theorem. In Theorem 3.1 we used a special abbreviation for a finite product of integers. These abbreviations occur frequently throughout probability theory.

Suppose $a_1, \cdots, a_n$ is a list of numbers. The sum $a_1 + \cdots + a_n$ may be abbreviated $\sum_{k=1}^{n} a_k$. The numbers 1 and n are called the lower and upper limits, respectively, of the sum. They vary from one sum to another. The letter k is called a dummy variable. The interpretation is this: substitute $k = 1, 2, \cdots, n$ into a_k and then form the sum of the terms so obtained. Any letter may be used in place of k. Thus $\sum_{k=1}^{n} a_k$, $\sum_{i=1}^{n} a_i$, $\sum_{m=1}^{n} a_m$ are equivalent ways of writing the same sum.

Evaluate the following sums.

(*a*) $\sum_{k=1}^{n} k$. (*b*) $\sum_{k=0}^{n} k$. (*c*) $\sum_{i=3}^{7} i^2$. (*d*) $\sum_{i=2}^{5} i^{-1}$. (*e*) $\sum_{k=4}^{8}(i^2 - i + 2)$.

(*f*) $\sum_{i=0}^{3} \sin i$. (*g*) $\sum_{i=-4}^{0} 2^{i+1}$.

Sometimes the dummy variable does not appear in a sum, as in $\sum_{k=1}^{n} 3$. The interpretation is that for each k we assign the value 3, or whatever the expression is following the summation sign. Thus we are to add n threes in this example. The result is $3n$ for this sum. Sometimes the limits are omitted from a summation. In this case they are understood from context. Also, there are different ways of specifying the range of the dummy variable.

Evaluate the following.

(*a*) $\sum_{k=1}^{8} 5$. (*b*) $\sum_{0<n<m} n$. (*c*) $\sum_{i<8} \sqrt{i}$ (it is understood from context that i is positive, also). (*d*) $\sum_{i=-n}^{n} z$. (*e*) $\sum_{2<k<8} a_k$.

(*f*) $\sum_{i=1}^{n} [(i+1)^2 - i^2]$. (*g*) $\sum_{i=5}^{87} [\log (i+1) - \log i]$.

Sums like those in the last two examples, where the terms cancelled with the exception of parts of the first and last terms, are sometimes called telescoping sums and occur quite frequently.

The same discussion holds for finite products as for finite sums if we replace addition throughout by multiplication and replace Σ by Π. For instance,

$$\prod_{k=1}^{n} a_k = a_1 \times a_2 \times \cdots \times a_n$$

Evaluate the following.

(*a*) $\prod_{i=1}^{n} \frac{i+1}{i}$. (*b*) $\prod_{m=1}^{n} \frac{m^2}{m^2 - 1}$. (*c*) $\prod_{i=-80}^{1000} i^{45}$. (*d*) $\prod_{5<i<8} i^0$.

(*e*) $\prod_{k<7} \frac{k}{k^2+1}$, where k is understood from context to be positive.

(*f*) $\prod_{k=-3}^{3} 2$. (*g*) $\prod_{k=1}^{n} 2^k$.

Certain formal rules are often helpful in manipulating sums and products.

Theorem 1. $\Sigma(a_k + b_k) = \Sigma a_k + \Sigma b_k$, $\Sigma c a_k = c \Sigma a_k$, and $\Pi(a_k b_k) = \Pi a_k \Pi b_k$, where the limits are the same on both sides of each equation.

Proof. Informally, we observe that if we expand both sides of each equation, there is a one-to-one matching or correspondence between the terms on the two sides. Formally, we can use induction, discussed below.

Double and multiple sums and products have similar interpretations. For example, $\sum_{k=1}^{5} \sum_{j=0}^{8} a_{j,k}$ means the sum of the 45 terms

$$(a_{0,1} + \cdots + a_{8,1}) + (a_{0,2} + \cdots + a_{8,2}) + \cdots + (a_{0,5} + \cdots + a_{8,5})$$

obtained by substituting in all $45 = 9 \times 5$ possible pairs j, k, where j ranges from 0 to 8 and k ranges from 1 to 5. We set $k = 1$ and let j range from 0 to 8, then set $k = 2$, and continue until we have set $k = 5$ and let j range from 0 to 8. Observe that the order of double or multiple summation is immaterial. The expression $\sum_{j=0}^{8} \sum_{k=1}^{5} a_{j,k}$ is simply the same sum with the terms rearranged. (With double sums where the dummy variables can each have infinitely many values, it is no longer true that the sum must be the same when the order of summation is reversed.)

Transformations of the dummy variable of the form $m = k + b$, with b an integer, are useful in manipulation of sums. Consider $\sum_{k=4}^{n} (k - 4)$. We let $k - 4 = m$ both in the limits and in the expression being summed. The lower limit $k = 4$ is equivalent to $m + 4 = 4$ or $m = 0$. The upper limit $k = n$ becomes $m + 4 = n$ or $m = n - 4$. The expression being summed is simply m. Thus the sum transforms into $\sum_{m=0}^{n-4} m$. In combining sums, this device is often useful.

Example 1. Find $S = \sum_{j=3}^{n+3} j + \sum_{k=-5}^{n-5} (k - 1)$. Making the transformations $j - 3 = i$ and $k + 5 = m$, we obtain

$$S = \sum_{i=0}^{n} (i + 3) + \sum_{m=0}^{n} (m - 6).$$

If we replace i by m in the first sum and use Theorem 1, we have

$$S = \sum_{m=0}^{n} (m + 3) + \sum_{m=0}^{n} (m - 6) = \sum_{m=0}^{n} (2m - 3) = 2 \sum_{m=0}^{n} m - 3(n + 1)$$

Using the result of Example 2 which follows, we find that the last sum is $n(n + 1)/2$, so the final result is $n(n + 1) - 3(n + 1) = (n - 3)(n + 1)$.

The principle of mathematical induction is used to prove assertions that hold for an infinite collection of integers.

Principle of Mathematical Induction

For each integer $n \geq N$, let $P(n)$ be a proposition (statement or assertion). If $P(N)$ is true and, whenever $P(n)$ is true it follows that $P(n + 1)$ is true, then $P(n)$ is true for all $n \geq N$.

We take this as an axiom. It will not be proven from prior principles. Some illustrations of its use in proofs follow.

Example 2. Prove that $\sum_{k=1}^{n} k = \dfrac{n(n + 1)}{2}$. For $n = 1, 2, \cdots$, let $P(n)$ be the proposition that $\sum_{k=1}^{n} k = \dfrac{n(n + 1)}{2}$. $P(1)$ is true: the verification reduces to $1 = 1 \times 2/2$. Suppose $P(n)$ is true. We wish to establish $P(n + 1)$. We have

$$\sum_{k=1}^{n+1} k = \sum_{k=1}^{n} k + (n + 1) = \frac{n(n + 1)}{2} + (n + 1) = \frac{(n + 1)(n + 2)}{2}$$

This proves $P(n + 1)$. We used $P(n)$ to replace the second sum.

Example 3. Prove that $n^{n/2} \leq n! \leq n^{n-1}$ for all positive n, with strict inequality if and only if $n > 2$.

Let $P(n)$ be the proposition to be proven. $P(1)$ is obviously true. Suppose $P(n)$ is true. To establish $P(n + 1)$, we have for the right side $(n + 1)! = n!(n + 1) \leq n^{n-1}(n + 1) \leq (n + 1)^n$. Thus $n! \leq n^{n-1}$ for all n. This is intuitively obvious, of course, for we simply replace each factor $2, \cdots, n$ in $n!$ by n. It is also apparent from this that the inequality is strict if and only if $n > 2$. Alternately, observe that if the inequality is strict in $P(n)$

then the first inequality in the proof of $P(n + 1)$ is strict hence the inequality is strict for $P(n + 1)$. Now $1! = 1^{1-1}$ and $2! = 2^{2-1}$ but $3! < 3^2$. Thus inequality holds for $N = 3$, and whenever it holds for n it holds for $n + 1$. Thus by induction it holds for all $n \geq 3$. We have used induction with $N = 3$ here.

To establish the left side, note that "$n^{n/2} \leq n!$ with strict inequality if and only if $n > 2$" is equivalent to "$n^n \leq (n!)^2$ with strict inequality if and only if $n > 2$." To establish the latter, we first verify that $P(1)$ and $P(2)$ hold with equality and $P(3)$ holds with inequality. Suppose $P(n)$ is true. Then $[(n + 1)!]^2 = (n!)^2(n + 1)^2 \geq n^n(n + 1)^2$, and if $P(n)$ has strict inequality, so does this expression. This is greater than or equal to $(n + 1)^{n+1}$ if and only if $n \geq (1 + 1/n)^n$. We shall show later that $(1 + 1/n)^n \leq e = 2.71828 \cdots$, the base of natural logarithms. Thus for $n \geq 1$, $n! \leq n^{n-1}$. The equality is strict for $n > 2$ due to the strictness of the inequality, for $n > 2$, $(n!)^2(n + 1)^2 \geq n^n(n + 1)^2$.

Example 4. *The binomial theorem.* For $n \geq 0$, let $P(n)$ be the proposition

$$(x + y)^n = \sum_{k=0}^{n} b(n,k)x^{n-k}y^k,$$

for any pair of real numbers x and y, where 0^0 is interpreted as 1.

Remark. The expression 0^0 is undefined. If we allowed $x = 0$ or $y = 0$ in the theorem, such nonsensical terms would appear. Alternately, we could attempt to define 0^0, for purposes of this theorem, so that when $x = 0$ and $y = 0$, true statements result. How shall we define 0^0? Suppose $x = 0$ and $y = 1$. Then $P(n)$ reduces to $1 = x^0$, since $b(n,0) = 1$. Thus we must define $0^0 = 1$. It can now be verified that this choice makes the binomial theorem true.

Proof of the theorem. $P(0)$ is true. Suppose $P(n)$ is true. We establish $P(n + 1)$. We have

$$\begin{aligned}(x + y)^{n+1} &= (x + y)(x + y)^n = (x + y)\sum_{k=0}^{n} b(n,k)x^{n-k}y^k \\ &= \sum_{k=0}^{n} b(n,k)x^{n+1-k}y^k + \sum_{k=0}^{n} b(n,k)x^{n-k}y^{k+1}\end{aligned}$$

Make the change of dummy variable in the second sum of $m = k + 1$ and in the first sum of $m = k$. Combining the sums gives

$$x^{n+1} + \sum_{m=1}^{n} [b(n,m) + b(n,m - 1)]x^{n+1-m}y^m + y^{n+1}$$

Since

$$b(n,m) + b(n,m - 1) = b(n + 1, m)$$

can be established by induction, the result follows.

An important practical limitation when using induction is that the statements to be proven must be discovered by other means. Induction is only used to verify them. A second limitation is that the use of $P(n)$ does not always help to prove $P(n + 1)$.

Some proofs are easier when other equivalent forms of the induction principle are used. There are more complicated proofs which involve a "double induction" (on pairs of integers).

SUGGESTED READINGS

Birkhoff and MacLane, Chapter I, Sections 1–5, for a discussion of induction.

EXERCISES

4.1 Prove by induction that $\sum_{k=1}^{n} \log a_k = \log (\prod_{k=1}^{n} a_k)$.

4.2 Prove Theorem 1 by induction.

4.3 Prove by induction that $\sum_{k=1}^{n} k^2 = \dfrac{n(n+1)(2n+1)}{6}$.

CHAPTER 2

Probability Measures On Finite Sample Spaces

In Chapter 1 we studied experiments which had a finite number of equally likely elementary outcomes. We were able to assign and often compute probabilities in a natural way. We are going to abstract and generalize this—a powerful technique throughout mathematics—so that probabilities can be assigned and computed when the elementary outcomes are no longer "equally likely." This generalization is already necessary to describe the results of so simple an experiment as tossing a biased coin.

Three centuries of experience has culminated in the modern view in which we carefully scrutinize the concept of "outcome" by the use of the theory of sets, and in which we abstract the properties of our numerical assignment of probabilities, using the theory of sets and the idea of a function. This leads us to the concept of a probability measure. We begin with the theory of sets.

1. ELEMENTARY SET THEORY

The list of outcomes of an experiment is an example of what mathematicians call a set. The term "set" is generally taken to be one of the primitive or undefined terms of mathematics, just as "point" is one of the primitive terms of plane geometry. In other words, "set" is not defined by other mathematical terms, just as "point" is not defined by other geometrical terms.

But the term "set" has an intuitive meaning. It is taken to be synonymous with "collection," "aggregate," or "ensemble." Sometimes we use these terms as synonyms for set. There are also very diverse types of sets such as the set of all positive integers, the set of all equilateral triangles, the set of all ideas, the set of all elementary particles in the universe, the set of all blades of grass, etc. These examples show that sets are defined by specifying their "members" or "elements." The examples also show that sets may be ither finite or infinite. A **finite** set is one whose members can be comletely labeled by using each of the positive integers from 1 up to and icluding some integer N, once and only once. The finite set is then said) have N elements. An **infinite** set is one which cannot be so labeled. /hich of the examples are infinite sets?

The notion of sets is one of the most widespread and important in all of athematics.

It is customary to call the set of outcomes of an experiment a **sample ace**. There is no mathematical need for this distinction. It probably ises from the fact that much of probability and its applications is conrned with sampling. That is, certain members of a set are examined. le list of such members is called a "sample." Inferences are then made om the observed properties of the sample about properties of the entire t, or population.

For example, a number of individuals may be polled on the way they end to vote in a forthcoming election. Inferences may then be made as the outcome of the election. More dramatically, we have all watched mputers on television use probability methods to analyze scattered ly returns (the sample) in recent presidential elections. The computers n project (predict within certain limits of accuracy, which increase as sample size increases during the evening) the actual outcome of the ction, both as to popular and electoral votes.

't is useful when analyzing a sample to consider the set or "space" of all ssible samples. It is then natural to call this set or space of samples a mple space" and then to go one step further to call *all* sets of outcomes experiments, whether from sampling or not, "sample spaces."

n set theory one generally limits discussion to one "big" set, the versal set, and subsets thereof. A subset E is simply a set made up of e of the members of the universal set, S. In particular, the set S itself onsidered to be a subset. Any other subset is called a proper subset. A ial subset of great usefulness is the subset which contains no members. called the empty set and will be denoted by $\varnothing$. For us, a universal set ill always be understood. Therefore we will refer to subsets simply as s."

ı probability theory, S will be the set of all elementary outcomes of e prescribed experiment, that is, a sample space. The elementary omes of the experiment are the members or elements of S. Subsets of rrespond to events. The set S corresponds to the certain event "some ome occurs." The set $\varnothing$ corresponds to the impossible event "no ome occurs." We often indicate a set by curly brackets enclosing a plete or an indicated list of members. For example, {a,e,i,o,u,} is the f vowels. Sets are also indicated by $\{s \text{ in } E: P\}$ where E is a set and P

is a property which further restricts it. We use small letters like s to denote members of S and capital letters like E to denote subsets. The expression $\{s \text{ in } S: P\}$ is read "the set of (members) s in S such that property P holds." The word "in" is often replaced by "$\in$" when it designates membership in a set. For example, $\{2, 4, 6, 8, \cdots\}$ and $\{n: n = 2m\}$ each might stand for the set of positive even integers, provided that in the last expression it was understood that n and m represented positive integers.

Strictly speaking, there is a logical distinction between an element x of a set S and the subset $E = \{x\}$ of S which consists of one element x. We shall refer to the one point subsets of a sample space, each consisting of a single elementary outcome, as elementary events. For an S with n points we will discuss the elementary events (subsets) $E_1, \cdots, E_n$.

We now define several concepts by which we can study the relations between sets, and manufacture other sets. Each concept is followed by a list of useful observations and consequences. The reader should immediately convince himself of these with informal reasoning.

Definition 1. If A is a set and every member of A is also a member of B, then A is a subset of B, written $A \subset B$, or equivalently, $B \supset A$.

(*a*) *Reflexivity*. $A \subset A$ for every set A.
(*b*) *Antisymmetry*. $A = B$ if and only if both $A \subset B$ and $B \subset A$.
(*c*) *Transitivity*. If $A \subset B$ and $B \subset C$ then $A \subset C$.
(*d*) For every set A, $\varnothing \subset A \subset S$.

Theorem 1. A finite set with n elements has $b(n,k)$ distinct subsets with k elements, where $0 \le k \le n$, and has a total of 2^n subsets.

Proof. The first assertion is a restatement of Theorem 1.3.3. The second follows from the observation that a subset is determined by specifying whether each element is in the set or is not. Thus we have two choices for each of n elements, yielding 2^n possible results.

Alternately, we can observe that the number of subsets is, from the first assertion, equal to $\sum_{k=0}^{n} b(n,k)$. But by the binomial theorem,

$$2^n = (1+1)^n = \sum_{k=0}^{n} 1^{n-k}1^k b(n,k) = \sum_{k=0}^{n} b(n,k).$$

If S is any set, the new set whose elements are the various possible subsets of S is often written 2^S. It is called the **set of all subsets of** S.

Definition 2. The set A' such that x is in A' if and only if x is not in A is called the complement of A.

(*a*) $\varnothing' = S$ and $S' = \varnothing$.

Definition 3. If S is any set, the union $A \cup B$ of the sets A and B is the set of elements which are in at least one of the sets A or B. The union

$_{=1} A_i$ of the sets $A_1, \cdots, A_n$ is the set of elements which are in at least
: of the given sets.

a) $A \cup S = S$ for all A.
b) $A \cup \varnothing = A$ for all A.
c) If $A \subset B$, then $A \cup B = B$.
d) *Associativity*. $A \cup (B \cup C) = (A \cup B) \cup C$.
e) *Commutativity*. $A \cup B = B \cup A$.

We define the union of an empty collection of sets to be the set ϕ. This is
ısible: no element is in at least one set in the empty collection.

)efinition 4. The set $\bigcap_{i=1}^{n} A_i$, or $A_1 \cap \cdots \cap A_n$, is the set of elements
ich are in all of the sets $A_1, \cdots, A_n$. It is the intersection of the given
;. We frequently omit the $\cap$ and simply write $A_1 A_2 \cdots A_n$ for the
ɪrsection. We define the intersection of an empty collection of sets to be
set S.

Note that $'$ is a **unary** operation on sets, that is, it takes *one* set A and
›duces a new set A'. In contrast, $\cup$ and $\cap$ are each binary operations,
.t is, they take pairs of sets A and B and produce new sets $A \cup B$ and
$\cap B$, respectively. If $*$ is any binary operation which satisfies the
ociative law $a * (b * c) = (a * b) * c$, we say it is associative. If it
isfies the **commutative** law $a * b = b * a$, we say it is commutative.
r example, if a, b, and c are real numbers, the binary operations of
lition and multiplication are each associative and commutative. The
der should verify that subtraction is neither.

)efinition 5. Two sets A and B are disjoint if $A \cap B = \varnothing$. When A
l B are disjoint, we sometimes write $A + B$ for $A \cup B$. A collection
$, \cdots, A_n\}$ of sets is disjoint if $A_i \cap A_j = \varnothing$ for each distinct pair of
ices i and j. A disjoint collection of sets is called a subpartition. A
partition $\{A_1, \cdots, A_n\}$ such that $\bigcup_{i=1}^{n} A_i = S$ is a partition of S. Any
ection of sets $\{B_1, \cdots, B_n\}$ such that $\bigcup_{i=1}^{n} B_i = S$ is a cover or
ering of S.

Ve can represent sets and the relations between them by schematic
ȝrams like Figure 1. The rectangle represents the universal set S. In the
rectangle, the horizontally shaded circle is the set A and the vertically
ded circle is B. Elements of A or B can be represented by points (dots)
A or B, respectively. The complement of a subset is the part of the
angle exterior to the set. The union of A and B is the total shaded area
he left rectangle. The intersection of A and B is the crosshatched area.
: right rectangle shows how the sets A and B divide S into four subsets.
uch figures are called Venn diagrams. We see that a cover of S by sets
rom the point of view of a Venn diagram, precisely what we would

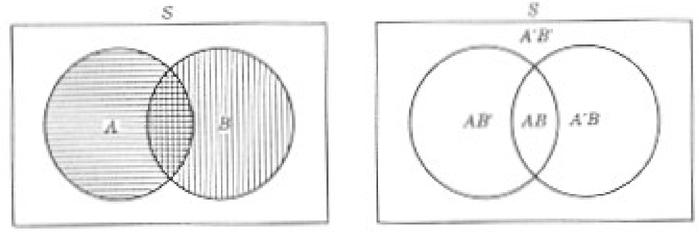


Figure 1. Illustrating relationships between sets for the general case of two sets A and B. The left diagram shows the sets A and B. The right diagram shows how they divide S into four subsets.

expect a "cover" to be intuitively. A partition divides S neatly into nonoverlapping pieces. Disjoint sets are simply ones which do not overlap.

Set theoretic expressions can be simplified with the aid of Venn diagrams.

Example 1. To find a simple expression for $(A \cup B)(A \cup B')$, we draw a Venn diagram for A and B as in Figure 2. We shade $A \cup B$ horizontally and $A \cup B'$ vertically. The intersection of the two shaded sets, the crosshatched area A, is the answer.

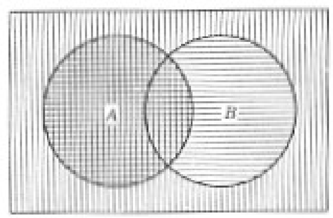


Figure 2. Simplifying $(A \cup B)(A \cup B')$ to A through the use of a Venn diagram.

Venn diagrams for three sets are also useful. For more than three sets they may become impractical.

Example 2. To check the equation $ABC = AB(B \cup C)$, we construct two identical diagrams and determine the (shaded) areas which correspond to each side of the equation. The equation is true if and only if the shaded areas are identical. This is illustrated by Figure 3, where we find ABC is a proper subset of $AB(B \cup C)$, in general.

Definition 6. The set $A - B = AB'$ is called the difference of A and B or the complement of B in (or relative to) A.

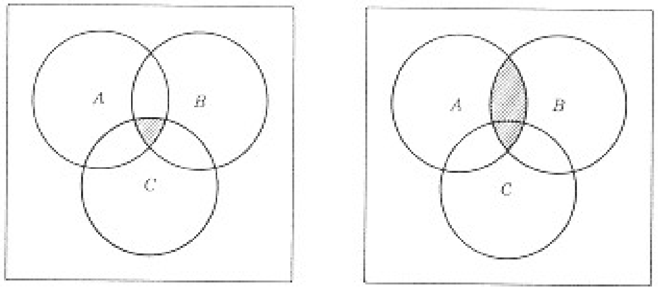


Figure 3. Showing that the equation $ABC = AB(B \cup C)$ is false, through the use of Venn diagrams. The shaded area in the left rectangle is ABC and the shaded area in the right rectangle is $AB(B \cup C)$.

The set theoretic operations of $'$, $\cup$, $\cap$, and $-$ have natural and fairly obvious interpretations in terms of events. If A is any event, then A' is the event "A does not occur." The expression $\bigcup_{i=1}^n A_i$ stands for the event "at least one of the events $A_1, \cdots, A_n$ occurs." The expression $\bigcap_{i=1}^n A_i$ stands for the event "all of the events $A_1, \cdots, A_n$ occur." The expression $A - B$ is the event "A occurs but B does not."

It is often desirable to use algebraic techniques, rather than pictorial techniques, in working with set theoretic expressions. In particular, Venn diagrams may be unwieldy when more than three sets are involved. The identities in the next two theorems are indispensable for algebraic manipulations with sets.

Theorem 2. *Distributive laws.*

$$A \cap \left(\bigcup_{i=1}^n A_i\right) = \bigcup_{i=1}^n (A \cap A_i) \quad \text{and} \quad A \cup \left(\bigcap_{i=1}^n A_i\right) = \bigcap_{i=1}^n (A \cup A_i)$$

Proof. To establish the first identity, suppose that the left side L is not empty and that x is a member of L. Then x is in A and also in at least one A_i. Let the first such i be i_0. Then x is in AA_{i_0} so it is a member of the right side R. Since x is an arbitrary member of L, it follows that $L \subset R$. If instead $L = \varnothing$, then $L \subset R$ trivially. Now suppose $R \neq \varnothing$ and let y be a member of R. Then for some i_0, y is a member of AA_{i_0}, so y is a member of both A and A_{i_0}, thus of both A and $\bigcup_{i=1}^n A_i$, and hence of L. Therefore $R \subset L$. If instead $R = \varnothing$, then $R \subset L$ trivially. Since $L \subset R$ and $R \subset L$, we have $R = L$.

Similar reasoning establishes the second assertion. (See Exercise 8.)

Theorem 3. *de Morgan's laws.*

$$\left(\bigcup_{i=1}^n A_i\right)' = \bigcap_{i=1}^n A_i' \quad \text{and} \quad \left(\bigcap_{i=1}^n A_i\right)' = \bigcup_{i=1}^n A_i'$$

Proof. Exercise 9.

Example 3. To find a simple expression for $(A \cup B)(A \cup B')$ by algebraic techniques, we have

$$(A \cup B)(A \cup B') = (AA) \cup (AB') \cup (BA) \cup (BB')$$

by repeated application of the first distributive law. This reduces to $A \cup (AB') \cup (AB) \cup \varnothing$. Since the last three terms are subsets of the first term A, this is simply A.

Example 4. To check the equation $ABC = AB(B \cup C)$, we note by the first distributive law that the right side is $(AB) \cup (ABC)$ or simply AB. It is easy to construct examples such that $AB \supsetneq ABC$ and hence the equation fails. Thus the identity is false in general.

There is a standard procedure for checking set theoretic or Boolean* equations. It depends on the fact that any expression involving n sets $A_1, \cdots, A_n$ and the Boolean operations $'$, $\cup$, and $\cap$, can be reduced in a finite number of steps to one of two canonical, or standard, Boolean forms. The first form is $\cup((A_1 \text{ or } A_1') \cap \cdots \cap (A_n \text{ or } A_n'))$. The union is over some of the 2^n possible indicated expressions. Since there are thus 2^{2^n} ways to pick the terms over which the union will be taken, there are 2^{2^n} distinct canonical Boolean forms. Two of these canonical Boolean expressions are identically equal, that is, equal for all particular collections of sets that might be considered, if and only if they have the same terms (see Exercise 10). In case each of the 2^n expressions is omitted, the first form is the union of the empty collection which is ϕ.

Note that the terms of a canonical form of the first kind correspond to the regions of a Venn diagram (Figure 1, right rectangle, for the case $n = 2$). To compare terms of two expressions is the algebraic equivalent of comparing shaded regions of Venn diagrams (Figure 3). To solve a problem like Example 4 algebraically we simply convert each side into canonical Boolean form.

The second canonical Boolean form is $\cap((A_1 \text{ or } A_1') \cup \cdots \cup (A_n \text{ or } A_n'))$ with the same qualifications as for the first form. It is obtained from the first form by applying one of de Morgan's laws. Conversely, the other de Morgan law allows one to obtain the first form from the second. Thus the two forms are equivalent. Sometimes it is easier to convert two expressions which are being compared into the first form and sometimes it is easier to convert them both into the second form. If each of the 2^n expressions is omitted, then the second form is the intersection over the empty collection, which is S.

* After the British mathematician George Boole (1815–1864) who founded the algebraic theory of sets.

SUGGESTED READINGS

Bell, Chapter 23, Boole.

Birkhoff and MacLane, Chapter XI, Sections 1 and 2, Boolean algebra.

Feller, Chapter 1, Sections 1–4 and selected problems from Section 8.

Goldberg, Chapter 1, sets.

Kemeny, Schleifer, Snell, and Thompson, Chapter II, Section 5, Voting coalitions, and Chapter III, Section 9, Voting power. Elementary set theory is used to study the significance of voting coalitions in voting bodies and to measure the power of subsets. The discussion is also available in Kemeny, Snell, and Thompson, Chapter II, Section 6—"Voting Coalitions," and Chapter III, Section 8—"Voting Power."

EXERCISES

1.1 Suppose A and B are subsets of a universal set S with 23 points, or elements. If A has 14 elements and B has 12 elements, what are the greatest and least numbers of elements that each of the following sets can possibly have? In each case, label the parts of a Venn diagram like that in Figure 2 to show a distribution of points which achieves the minimum or maximum you give.

(*a*) AB. (*b*) AB'. (*c*) $A'B$. (*d*) $A'B'$. (*e*) $A \cup B$.

1.2 Let S be the set of all wavelengths of visible light.

(*a*) Consider two filters (pieces of colored glass). If A is the set of wavelengths which filter 1 passes and B is the set of wavelengths which filter 2 passes, what is the set of wavelengths that is passed when both filters are used?

(*b*) Consider two pigments. If A is the set of wavelengths which pigment 1 reflects and B is the set of wavelengths which pigment B reflects, what set of wavelengths is reflected when the two pigments are mixed? (Assume the pigments are unchanged after the mixing.)

1.3 Let $S = \{1,2,3\}$. How many subsets of S are there? List them.

1.4 The number of atoms in the known universe has been estimated as 10^{79}. How many elements must a set S have to guarantee that it has at least this many subsets?

1.5 Simplify the following expressions, both by using Venn diagrams and by algebraic methods. (Strive for brevity and elegance.)

(*a*) $(A \cup B)B'$

(*b*) $(A \cup A')(B \cup B')A$

(*c*) $((AB') \cup (C \cup B'))(A' \cup C)'$

1.6 Write each of the following in each of the two canonical Boolean forms.

(*a*) $A \cup B$

(*b*) AB'

(*c*) $((AB) \cup C')(A' \cup C)$

(*d*) $((A \cup B)(C'D)) \cup (BD')'$

1.7 Which of the following equations are true?

(*a*) $A \cup B = (A - B) \cup (B - A) \cup (AB)$

(*b*) $A \cup B \cup C = (A - (B \cup C)) + (B - (A \cup C)) + (C - (A \cup B)) + (AB - C) + (BC - A) + (CA - B) + ABC.$

(*c*) $(AB) \cup C = (A'B'C) \cup (AB'C')$

1.8 Prove that $A \cup (\bigcap_{i=1}^{n} A_i) = \bigcap_{i=1}^{n} (A \cup A_i)$.

1.9 Prove de Morgan's laws.

1.10 Prove that two canonical Boolean forms of the first kind are identical if and only if they have the same terms. If desired, the same result now follows for the second form by de Morgan's laws.

***1.11** Let $a * b$, where a and b are positive integers, represent exponentiation, that is, $a * b = a^b$. Show that $*$ is not commutative. Is $*$ associative? Note that since $*$ is not commutative, there are now two distributive laws for $*$ with respect to addition. The **left distributive law** asserts that $a * (b + c) = (a * b) + (a * c)$. The **right distributive law** asserts that $(b + c) * a = (b * a) + (c * a)$. Is either one true for the given situation?

***1.12** Using the fact that the commutative law holds for $\cup$, show that there are at least $n!/2$ formally different ways of computing the expression $\bigcup_{i=1}^{n} A_i$, $n \geq 2$.

****1.13** Three red hats and two black hats are shuffled in an urn. Three hats are drawn and put on the heads of three blindfolded men. When the blindfolds are removed, each man can see which color hat the others are wearing but not the color of his own hat. Assume that the men cannot communicate, that they know the hats were selected from three red and two black, and that they make full use of their knowledge of the colors of the hats which they see the other men wearing. Suppose that the men are asked in turn what color hat they are wearing and each can hear the answers given.

(*a*) Prove that one man can deduce what color hat he is wearing.

(*b*) Show that the problem is true for the case of n men, n red hats and $n - 1$ black hats, when $n \geq 2$.

(*c*) Show that for the case of n men, the assertion of the problem is false when the number r of red hats and b of black hats both are greater than or equal to n. Show that the assertion is true whenever either b or r is less than n.

2. FUNCTIONS

We have developed precise concepts to replace our informal notions about outcomes of experiments, or events. The next step in extending our probability theory is to develop precise concepts about the way in which we assign numerical probabilities to events.

We begin by studying the notion of "assignment" or "correspondence" itself. In mathematics this notion goes by the name of "function." It is one of the most elementary and far reaching of all mathematical concepts.

Definition 1. Let X and Y be sets. A function f from X to Y is a rule of correspondence which assigns to each of some (but not necessarily all) of the elements x of X an element $f(x)$ of Y. The subset $D \subset X$ to which f assigns elements of Y is the domain of f. The set X is called the domain space of f. The subset $R \subset Y$ of the elements so assigned is the range of f. The set Y is called the range space of f. To specify the function we must not only specify the rule of correspondence but D, X, R, and Y as well. We write $f\colon D \subset X \to R \subset Y$. The function f is said to map from X into Y.

When the domain and range are understood (as is generally the case), we simply write $f: X \to Y$. When X and Y are also understood, we may write f. If we wish to emphasize that f is a function, we may write $f(\cdot)$. When we write "$f(x)$," we say that we evaluate f at x. If $D = \phi$ so that f is defined nowhere, it is convenient to call f the empty function. This could arise if f were defined by an illegal expression, say $f(x) = x - 2 +$ or $f(x) = 2/0$, with Y the real numbers. It also arises in studying the composition $g(f(x))$ of two functions: when $R_f \cap D_g = \phi$, $g(f(x))$ is the empty function.

Example 1. Let X be the set of integers $\{0, \pm 1, \pm 2, \cdots\}$ and let Y be the real numbers. Let D be the non-negative integers and let $f(x) = x^2$ for each x in D. Then R is the set of squares of non-negative integers. If we replace D by the larger set X, we get a new function f^* which agrees with D but is defined for more (in fact, all) x in X. Such a function f^* is called an *extension* of f. Suppose we were to replace Y by some other set, for example, the set of integers J. Then the function g from $D \subset X$ to J defined by $g(x) = x^2$ is very "similar." But technically, it must be considered a different function than f since Y has been altered (to J).

Note that functions are **single-valued,** that is, to each x in X they assign at the most one $f(x)$ in Y. This is the same as saying that whenever $f(x_1) \neq f(x_2)$, then $x_1 \neq x_2$. (Different points in the range come from different points in the domain.) "Functions" which are not single-valued are called **many-valued.** They are of considerable mathematical importance but will not be needed by us.

Definition 2. A function f is onto if $R = Y$. It is one-to-one (1-1) if for each y in R there is precisely one x in D such that $f(x) = y$. Otherwise f is many-to-one. A 1-1 onto function with $D = X$ is biunique. If $f(x) = y_0$ for all x in X, where y_0 is the same for all x, then f is a constant function.

(*a*) f is 1-1 *if* and only if whenever $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$.

(*b*) In the theory of sets, two sets X and Y (finite or infinite) are said to have the same number of points if and only if there is a biunique correspondence between them.

(*c*) If X has at least two points, constant functions are many-to-one.

Example 2. Let $X = Y = \{1,2, \cdots\}$. If $f(x) = x + 1$, f is 1-1 and $D = X$ but f is not onto. If the range of f is finite, and $D = X$, then f must necessarily be many-to-one.

Example 3. If an experiment has a finite number N of equally likely elementary outcomes, the probability $P(E)$ of any event E is M/N, where M is the number of points in E. This rule P for assigning a probability to each subset of S is a function $P: 2^S \to Y$ where Y is the set of real numbers (this is not the only choice for Y but it is the customary one) and the domain of P is all of 2^S. In this example, the range of P is the set of $N + 1$ fractions $\{0, 1/N, 2/N, \cdots, (N-1)/N, 1\}$.

Definition 3. The characteristic function C_E of $E \subset S$ is the function from S to the real numbers defined by $C_E(x) = 1$ if x is in E and $C_E(x) = 0$ if x is not in E.

The reader should verify the following properties of characteristic functions.

(*a*) $C_{E'} = 1 - C_E$ (note that two functions are equal if and only if they are equal when evaluated at each x in S.

(*b*) $C_{A \cup B} = \max(C_A, C_B)$ and similarly for n sets $A_1, \cdots, A_n$. Given a finite set $(r_1, \cdots, r_n)$ of real numbers, we define $\mathbf{max}(r_1, \cdots, r_n)$ to be the greatest number in the set. If $(f_1, \cdots, f_n)$ is a finite set of functions defined on the same domain D of a set X into the real numbers, $\mathbf{max}(f_1, \cdots, f_n)$ is the function which has the value $\max(f_1(x), \cdots, f_n(x))$ at each x in D.

(*c*) $C_{A \cap B} = C_A C_B = \min(C_A, C_B)$ and similarly for n sets $A_1, \cdots, A_n$. The **product** fg **of two real-valued functions** f and g defined on the same domain D in a set X is the function $h = fg$ such that $h(x) = f(x)g(x)$ for each x in D.

(*d*) $C_{A-B} = C_A(1 - C_B) = C_A - C_A C_B$.

Characteristic functions are an important tool in many parts of mathematics. The correspondence (*a*) to (*d*) between the set theoretic operations on sets and the algebraic operations on the characteristic functions of those sets is responsible for much of their utility in probability theory.

EXERCISES

2.1 If $X = \{1,2\}$ has 2 points and $Y = \{1\}$ has 1 point, list all the functions from X to Y. It is convenient in this case to describe a function f by listing the pairs $(x, f(x))$ for each x in D.

2.2 If $X = \{1, \cdots, n\}$ and $Y = \{1\}$, how many functions are there from X to Y?

2.3 If $X = \{1,2\}$ and $Y = \{1,2\}$, list the functions from X to Y. How many are there? Which are 1-1? Which are onto?

2.4 If $X = \{1, \cdots, m\}$ and $Y = \{1, \cdots, n\}$, how many functions are there from X to Y? Verify that your formula gives the correct answer in the special cases previously solved.

2.5 Give (your own) examples of functions such that $D = X = Y =$ real number system and:

(*a*) f is 1-1 and onto.
(*b*) f is not 1-1 and not onto.
(*c*) f is 1-1 and not onto.
(*d*) f is not 1-1 but is onto.
(*e*) Do (*a*) to (*d*) when D is the subset of positive numbers.
(*f*) Do (*a*) to (*d*) when D is the set of numbers $\{x: 0 \leq x \leq 1\}$.

3. PROBABILITY MEASURES ON FINITE SAMPLE SPACES

The function P of Example 2.3 "measures" the probability of each event for an experiment with N elementary equally likely outcomes, $E_1, \cdots, E_N$. Notice that each *elementary* outcome E_i is assigned the probability $1/N$ and that $\sum_{i=1}^{n} P(E_i) = 1$. In the case of a fair coin, $P(H) = P(T) = \frac{1}{2}$. It seems natural, in the case of a biased coin, for example, to assign probabilities $P(H) = p$ to heads and $P(T) = q$ to tails, where p and q are now unequal in general. We expect $p + q = 1$, hence $q = 1 - p$. If the coin were two-headed, for example, we choose $p = 1$, $q = 0$. If the coin were weighted on the heads side, some choice where $p > \frac{1}{2}$ would seem appropriate. We will see later how to arrive at an appropriate value.

In the general case of N elementary outcomes $E_1, \cdots, E_n$, which are no longer equally likely, let $P(E_i) = p_i \geq 0$ with $\sum_{i=1}^{n} p_i = 1$. Define $P(E) = \Sigma P(E_i)$, with the sum taken over those $E_i \subset E$, to be the probability of an event $E \subset S$. The probability function P is a **probability measure** on the finite sample space S. It is easy to verify that any such P obeys the following rules.

P1. *Normality:* $P(S) = 1$.

P2. *Non-negativity:* $P(E) \geq 0$ for all $E \subset S$.

P3. *Additivity:* $P(E + F) = P(E) + P(F)$ for disjoint sets E and F.

There are many other rules which P satisfies but they can all be deduced from $P1$ to $P3$. The reader should so deduce the following.

(*a*) $P(\varnothing) = 0$.

(*b*) If $E \subset F$ then $P(E) \leq P(F)$. Thus $0 \leq P(E) \leq 1$ for all $E \subset S$.

(*c*) $P(E') = 1 - P(E)$.

(*d*) $P(E \cup F) \leq P(E) + P(F)$.

(*e*) $P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$.

A formal proof of (*e*) uses a mathematical induction. If the sets $A_1, \cdots, A_n$ are disjoint we can similarly prove that $P3$ generalizes to $P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$.

The rules $P1$ to $P3$ yield the following definitions.

Definition 1. If S is a finite set and P is real-valued function defined on all of 2^S which satisfies $P1$ to $P3$, then P is a probability measure for the finite sample space S. For the special case where each elementary outcome is equally likely, P is called the equiprobable measure.

Once we have specified $P(E_i) = p_i$ for the elementary outcomes, the probability measure is determined by $P3$ for all events E. Therefore, in the future we will simply specify it for the E_i, rather than for all E.

It is useful to think of P as a mass function which assigns mass $\{\Sigma p_i : E_i \subset E\} = P(E)$ to a set E so that the mass of each set E is the sum of the masses of the "points" which compose it. The whole space has mass $P(S) = 1$. This picture continues to be useful in continuous probability (Ch. 4).

If S is a set with one point, there is only one way to assign a probability measure, namely $P(S) = 1$, $P(\varnothing) = 0$. But if S has two or more points, there are an infinite number of ways to assign a probability measure. Suppose for instance that $S = \{s_1, s_2\}$ has two points, corresponding to two elementary outcomes $E_1 = \{s_1\}$, $E_2 = \{s_2\}$. Then any choice of p_1 and p_2 such that $0 \leq p_1 \leq 1$ and $p_2 = 1 - p_1$ yields a probability measure. Conversely, all probability measures are obtained in this way. Therefore, if we have a probability situation, it is not enough to specify the sample space S. We must also describe the probability measure.

Definition 2. A finite probability space is a pair (S,P) where S is a finite set and P is a probability measure on S.

Example 1. Two coins are tossed. Castor argues that the probability of the event E that both coins will show heads is $1/3$ because there are three outcomes: no head, one head, and two heads. There is one way for E to occur out of three possible ways so $P(E) = 1/3$. His twin Pollux argues that the probability is $1/4$, because there are four outcomes: (H,H), (H,T), (T,H), and (T,T), and E occurs in only one way. Who is right?

Castor and Pollux are simply assigning different probability measures to the same sample space S. We take the point of view here that Castor is lumping (H,T) and (T,H) together and making any assignment for $a = P((H,T))$ and $b = P((T,H))$ such that $a + b = 1/3$ and a and b are both non-negative. The question reduces to determining which probability measure is correct.

If the coins are true, then by assumption Pollux is right. It can happen that both are wrong. This will generally be the case when at least one of the coins is unfair and they are tossed "independently." In particular, if the coins are tossed "independently" of one another, Castor must be wrong* as we will see in Exercise 7.5 whereas Pollux can be right.

It is generally true that for a given experiment, many choices for the sample space and the probability measure are equally valid. Of course, some choices are more natural or more convenient than others.

Example 2. A line is drawn on one edge of a coin, perpendicular to the faces. In addition to which side of the coin faces up, we can determine in which of the four compass directions the line most nearly lies. Thus we have eight outcomes, which we write as $S = \{HN, HE, HS, HW, TN, TE, TS, TW\}$. If the coin were true, we might argue that if in addition the coin is "fairly" tossed, then each of the directions is equally likely, so that the equiprobable measure would be appropriate. Call this probability space (S,P). Then we have $P(H) = P(T) = 1/2$.

* This is not surprising. Castor and Pollux are the famous Gemini twins of mythology. Pollux was immortal but Castor was a mere mortal.

This is consistent with our simpler description of a coin toss by the sample space $T = \{H,T\}$ and the equiprobable measure Q for T. Note that (S,P) is just a finer or more detailed description of the situation than (T,Q). Elementary events in T become compound events in S, but their probabilities measured by Q are the same when measured by P. If we are not interested in anything but the events H,T, then it is superfluous to consider a more refined description like (S,P).

EXERCISES

1 What sort of sample space and probability measure would you use to describe each of the following. A *qualitative* description of the probability measure is all that is expected.

(*a*) A loaded die is rolled and the uppermost face is recorded.

(*b*) A biased roulette wheel is spun and the resulting number is recorded.

(*c*) A package of a certain breakfast cereal is supposed to contain 14 ounces net weight. A package is selected "at random" and the contents are weighed to the nearest 0.1 ounce.

(*d*) A mortar is locked in position and accurately aimed at a certain point. A dummy shell is fired and the final distance from the target is recorded to the nearest yard.

(*e*) A sample of size N is taken from the output of a machine that rapidly produces many small parts (buttons, nails, stamps, coins, staples, etc.). The number k of defectives is recorded.

3.2 Consider a roulette wheel with 38 equally likely outcomes, 0, 00, 1, 2, . . . , 36. The number 0 and 00 are green. Eighteen of the numbers from 1 to 36 are black and eighteen are red. Similarly, the numbers 1, . . . , 36 are divided into the two equal groups "high" and "low," the two equal groups "odd" and "even," and into groups of a dozen 1 to 12, 13 to 24, 25 to 36. Still other divisions are practiced. Describe the probability measure when the sample space consists of the following elementary events.

(*a*) 00, 0, 1, . . . , 36.

(*b*) Green, red, black

(*c*) Green, odd, even

(*d*) Green, low, high

(*e*) Green, first dozen, second dozen, third dozen

(*f*) 00, 0, odd and red, odd and black, even and red, even and black

Which of the probability spaces from (*a*) to (*f*) are refinements of which others?

3.3 *Yarborough.* Aces, Kings, Queens, Jacks, and Tens are referred to in bridge as honors. The Duke of Yarborough (a century or so ago) is reported to have regularly wagered 1000 pounds to 1 pound that a player would be dealt a hand containing at least one honor. A hand without an honor is now commonly referred to as "Yarborough." What is the probability that a hand will be a Yarborough? Did the Duke have a good bet?

3.4 What is the probability that a randomly dealt hand at bridge contains no Ace or face card (King, Queen, or Jack).

3.5 Refer to Example 2. Give a probability space (U,R) which is more detailed than (T,Q) but which is "incompatible with" or "not comparable to" (S,P), that is, neither (S,P) or (U,R) is a refinement of the other.

3.6 An urn contains $m > 0$ black balls and $n > 0$ white balls. A random sample of two balls is drawn without replacement. Let W be the event that both are white, M the event that one is white and one is black, and B the event that both are black. Compute $P(B)$, $P(W)$, and $P(M)$. Verify that

$$P(B) + P(M) + P(W) = 1$$

4. PROBABILITY MEASURES AND RELATIVE FREQUENCIES

In the equally likely case, we had a natural procedure for determining the $P(E_i)$. But when the elementary outcomes are not equally likely, it happens commonly that the numbers $P(E_i)$ cannot be deduced from a priori reasoning. Yet finding the values of these numbers may be of primary interest. One procedure for doing this is to repeat the experiment many times and to keep track of how often each of the various elementary outcomes occurs. If in n experiments or **trials** each E_i occurs n_i times, it seems plausible to estimate $P(E_i)$ by $P(E_i) = n_i/n$. These numbers n_i/n are called the **relative frequencies** of occurrence of the events E_i.

The numbers n_i/n are only approximations of some sort to hypothesized "true" values. Suppose, for instance, that only one trial is made. Then the estimates are 0 for all but one $P(E_i)$ and 1 for the remaining $P(E_i)$. More generally, if n trials are made with n less than the number N of elementary outcomes, some $P(E_i)$ will be estimated incorrectly as 0, provided all $P(E_i)$ have a nonzero chance of occurring. Relative frequencies are always rational numbers, that is the quotient of two integers, so an irrational $P(E_i)$ can never be estimated exactly.

In the equally likely outcomes case, where the $P(E_i)$ are known, experiments show that as the number of trials increases, n_i/n "tends to" (or "approaches") $P(E_i)$ in a sense which will later be made precise. When the $P(E_i)$ are not known in advance, the numbers n_i/n are still found to tend to certain values as n becomes large. We take these values as the correct ones for the $P(E_i)$. We can only estimate the correct values in this way, but by taking n sufficiently large, we can make the uncertainty in the estimate as small as we like.

Instead of assuming that the $P(E)$ are given for all events E, suppose that we assume they are approximately determined as relative frequencies by repeated trials. What rules do the $P(E)$ satisfy? For any event E, $P(E) = M/N$ where M is the number of times an event E was observed in N trials. The approximation improves as N increases. This point of view towards probabilities—the relative frequency view—is particularly favored by many (perhaps most) statisticians. Is the relative frequency view of probability consistent with the one inspired by consideration of equally likely alternatives? We check the rules in Definition 3.1.

P1. First, $P(S)$ is always estimated by $N/N = 1$ so $P(S) = 1$ must be the number that the relative frequency estimates "tend to" as N increases.

P2. Since $M/N \geq 0$ always, the M/N must always "tend to" a number which is non-negative. (If, instead, the M/N got close to a negative number, then M/N would have to be negative sometimes, which is impossible.)

P3. The equation $P(E + F) = P(E) + P(F)$ is easily seen to be true when the values of $P(E)$, $P(F)$ and $P(E + F)$ are replaced by their relative frequency estimates. Since the equation therefore holds no matter how accurately the individual terms are approximated, it must hold when the true values are inserted. (This verbal argument can be replaced by a precise mathematical proof when we later have an exact definition of "tends to.")

Thus, the probabilities assigned by the relative frequency approach agree with our earlier Definition 3.1.

The relative frequency approach gives us an experimental method for determining values for the probabilities of events. We simply repeat an experiment "enough" times so that the experimental values for the probabilities of the elementary outcomes are as close as desired to the unknown limiting or "true" values. We shall see in Chapter 3 how to determine the number of repetitions of the experiment that are required for a given accuracy.

Example 1. *Roulette*. A roulette wheel has n pockets. In Nevada, there are generally 38 pockets, numbered $1, \cdots, 36, 0, 00$. In Europe, 00 is absent and there are only 37 pockets. The pockets are intended to be of equal size and the machine is designed so that when the ball is spun, it has equal probabilities of falling into each of the n pockets. A wheel in which the n outcomes are equiprobable is **true,** as opposed to **biased.** Wheels are initially well-machined and carefully balanced to make them true. However, they sometimes become sufficiently biased so that the player who is aware of this has an advantage over the operators. Bias can be detected by recording enough trials so that the relative frequencies of the various numbers can be estimated accurately. A statistician can often detect significant bias long before it becomes apparent to the operators of the game.*

Example 2. Suppose a large horizontal disc is carefully machined and balanced, like the rotor of a roulette wheel. Given a finite probability space with probabilities $p_1, \cdots, p_n$ for the elementary events, draw rays or "spokes" from the center of the disc to the edge, so that they divide the disc into angular sectors proportional to the p_i. The ith sector measures $p_i/2\pi$ radians. Now erect a thin horizontal pointer over the edge of the disc and pointing towards its center. When the disc is spun, the probabilities of the pointer being in any sector when it comes to rest (an elementary event) ought to be p_i. If the pointer is on the line, we make the convention that it points to the left of the two sectors. This is to eliminate any effects from the finite width of the lines. An experiment which records relative frequencies will then show that the probabilities of the various elementary events are as specified to high accuracy.

This example shows that any finite probability space can be realized in practice. Of course there are also many other ways to accomplish this.

* An account of actual casino experiences with detecting and exploiting biased roulette wheels, and some of the mathematics involved, is given by Allan Wilson, *The Casino Gambler's Guide*, Harper and Row, 1965.

SUGGESTED READINGS

Parzen, Emanuel, *Modern Probability Theory and Its Applications.* Chapter I and Chapter II, Sections 1 and 2, review the material up to this point, with probability initially motivated by relative frequencies. The exercises are a useful supplement.

Goldberg, Chapter 2, Sections 1–4.

EXERCISES

4.1 Toss a fair coin 100 times, recording the result. If h_n is the net number of heads at trial n, plot (n,h_n) on graph paper. Connect up these points and obtain a polygonal line, each segment of which has slope 0 or 1. Tossing the coin is tedious. A simpler procedure is to refer to a table of random numbers. Count even digits as heads and odd digits as tails, and use a string of 100 consecutive digits.

Draw a straight line connecting (0,0) and (100,50). This line has slope $\frac{1}{2}$, corresponding to the "true" probability of $\frac{1}{2}$ for heads on each trial. Note that the slope of the line joining (0,0) to (n,h_n) gives the observed relative frequency of heads at time n. Comparing the slope of this line with the slope of the previously drawn line makes it easy to see visually whether or not the observed frequency at a given time is "close" to the true probability.

Does the difference tend to get small as n increases? (See Feller, page 84, for a graph of the result of 10,000 tosses of a coin.)

If you have access to a computing machine, and suitable skill, you can readily simulate this experiment for a very large number of tosses. Program the machine to generate pseudorandom numbers. Then print out n and h_n/n, say, every 1000 or 10,000 tosses, for 100,000 or 1,000,000 tosses. Does the sequence of h_n/n values which are printed out seem to tend to $\frac{1}{2}$ as n increases?

4.2 We now use the idea in the preceding exercise to estimate an "unknown" probability. Refer to a table of random numbers. Consider these numbers in consecutive groups of 5. In each group, record a success H if 1 appears somewhere in the group and a failure T if 1 appears nowhere in the group. Do this for 100 groups and plot as in Exercise 4.1. Now draw in a straight line connecting (0,0) and $(100, h_{100})$. The slope $h_{100}/100$ of this line is your estimate of the true probability $P(H)$ of success.

Compute the true probability and compare with $h_{100}/100$. If convenient, simulate with a computing machine as above. The numbers $h_{1000}/1000$, $h_{2000}/2000, \ldots$ should tend towards the computed true probability.

4.3 Suppose k people are chosen "at random."

(*a*) What is the probability that some two people have the same birthday (day and month the same, the year may be different)? Assume 365 days in a year and that all assignments of birthdays to the n people are equally likely.

(*b*) If you are one of the k people, what is the probability that one of the others has your birthday?

4.4 Give an example in which the elementary outcomes for unordered sampling without replacement are equally likely even though they are not equally likely for ordered sampling without replacement. (See the remark after Theorem 1.3.3.)

4.5 Show that $P1$ to $P3$ are independent, that is if any one of the three conditions for a probability measure are dropped, then there are examples which satisfy the other two conditions but violate the omitted one. Can you find all possible examples in each of the three cases?

5. CONDITIONAL PROBABILITY AND INDEPENDENCE

We frequently wish to find the probability of an event given some condition or restriction. For example, we might wish to know the probability that an Ace will be drawn from a well-shuffled deck, given the condition that the deck contains no Kings. The condition or restriction simply limits the possible outcomes to some subset B of the sample space. Thus we wish to find $P(A \mid B)$, the probability that A occurs, given the additional information that B occurs. Such probabilities are known as **conditional probabilities.**

If the elementary outcomes are equally likely, suppose that there are n_A ways in which A can occur, n_B ways in which B can occur, and n_{AB} ways in which both can occur. Then, given the sole information that B occurs, the n_B ways are still equally likely, each now with probability $1/n_B$. Since B occurs, A can now occur in only n_{AB} ways. Thus $P(A \mid B)$ is n_{AB}/n_B. Notice that this equals

$$\frac{P(AB)}{P(B)} \quad \text{since} \quad \frac{n_{AB}/n}{n_B/n} = \frac{n_{AB}}{n_B}$$

The relative frequency viewpoint leads to the same results. Here n_A, n_B and n_{AB} are instead the number of occurrences of A, B and AB, respectively, in a "large" number n of trials.

These considerations lead us to the following.

Definition 1. The conditional probability $P(A \mid B)$ that an event A occurs, given the condition that B occurs, is $P(AB)/P(B)$ if $P(B) \neq 0$. If $P(B) = 0$, $P(A \mid B)$ is not defined.

We read $P(A \mid B)$ as "the probability of A given B."

Example 1. Two coins are placed heads up side by side on a flat surface. They are welded together. When they are tossed, only two of the four outcomes (H,H), (H,T), (T,H) and (T,T) are now possible, that is (H,H) and (T,T). Suppose that $P((H,H)) = p$ with $0 < p < 1$ and $P((T,T)) = 1 - p$. Let H_i be the event "the ith coin shows a head" and T_i the event "the ith coin shows a tail." Then $P(H_2 \mid H_1) = 1$ and $P(T_2 \mid H_1) = 0$, corresponding to the fact that the outcome of the first coin fully determines the outcome of the second.

Example 2. $n + 1$ urns are numbered from 0 to n. The ith urn contains i white balls and $n - i$ black balls. A ball is drawn at random (that is all balls have equal probability of being drawn) from the total of $n(n + 1)$ balls. In this particular case we might do this by selecting an urn at random while blindfolded, then choosing a ball at random from the urn that was selected. (See Exercise 5.14.) If the ball selected is white, what is the probability that it came from the ith urn?

Let W be the event "a white ball is selected" and let U_i be the event "the ball came from the ith urn." We want to find $P(U_i \mid W)$. The total number of white balls is $0 + 1 + \cdots + n = n(n + 1)/2$. This yields $P(U_i \mid W) = 2i/n(n + 1)$. In particular, $P(U_0 \mid W) = 0$, $P(U_i \mid W)$ increases as i increases, and $\sum_{i=0}^{n} P(U_i \mid W) = 1$, all in agreement with our intuition.

There is an important special case of conditional probability, in which the probability of A is not affected by the information that B occurred, that is $P(A \mid B) = P(A)$. The event A is said to be independent of the event B. Using Definition 1, we have $P(AB) = P(A)P(B)$ in this case. Note that if A is independent of B, for which we assumed $P(B) \neq 0$, and if we further assume $P(A) \neq 0$, then $P(B \mid A) = P(AB)/P(A) = P(B)$. Thus B is also independent of A. Hence we simply say **A and B are independent.**

We have imposed the restrictions $P(A) \neq 0$, $P(B) \neq 0$, in defining independence. These restrictions were inherited from the definition of conditional probability. If instead of $P(A \mid B) = P(A)$, we use the equivalent $P(AB) = P(A)P(B)$ in defining independence, then no quotient appears and we can drop the restrictions. If we do so, and it happens that either $P(A)$ or $P(B)$ is zero, then $P(AB)$ is zero as well and both sides are zero. Because we can drop the restrictions on $P(A)$ and $P(B)$, we shall adopt this "multiplicative" form of the definition of the independence of two events. More generally we have the following.

Definition 2. Given the probability space (S,P), a collection of two or more events $A_1, \cdots, A_n$ is pairwise independent with respect to (S,P) if $P(A_iA_j) = P(A_i)P(A_j)$ for every pair of distinct indices, i and j. The collection is mutually independent with respect to (S,P) if

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

for every subset $i_1 < i_2 < \cdots < i_k$ of distinct indices.

The reader should verify the following.

(*a*) A mutually independent collection is pairwise independent.

(*b*) The definitions of mutual and pairwise independence coincide when $n = 2$.

(*c*) A subcollection of a pairwise independent, or of a mutually independent, collection is respectively pairwise or mutually independent.

(*d*) If a collection is pairwise independent, or mutually independent, the new collection obtained by adding one or more events having probability 0 or probability 1 (for example, copies of $\varnothing$ or S) is also, respectively, pairwise or mutually independent. It therefore follows [with the aid of (*c*)] that a collection is respectively, pairwise or mutually independent if and only if the subcollection obtained by removing all sets having probability 0 or 1 is also respectively, pairwise or mutually independent.

(*e*) If $\{A_1, \cdots, A_n\}$ is a pairwise independent collection such that $0 < P(A_i) < 1$ for all i, then $A_i \not\subset A_j$ (where $E \not\subset F$ means E is not a subset of F) for each pair of distinct indices i and j. Pictorially, each A_i "sticks out of" each A_j, $j \neq i$.

(*f*) The definition is unchanged if we replace the ordered subsets of k distinct indices $i_1 < i_2 < \cdots < i_k$ (there are $b(n,k)$ such) by the unordered subsets of k distinct indices (there are $(n)_k$ such).

(*g*) Pairwise independence imposes $b(n,2)$ conditions. [Recall that we define $b(n,k) = 0$ for $n < k$.] Mutual independence imposes $2^n - n - 1$ conditions.

Since $2^n - n - 1 = \sum_{k=2}^{n} b(n,k) > b(n,2)$ when $n \geq 3$, we might expect mutual independence to be a strictly stronger condition than pairwise independence when $n \geq 3$.

Example 3. A pairwise independent collection need not be mutually independent. To see this, let $S = \{1,2,3,4\}$ with probability 1/4 for each of the elementary outcomes. Let $A_i = \{i,4\}$, $i = 1, 2, 3$. Then $A_iA_j = \{4\}$, $i \neq j$, and $P(A_iA_j) = 1/4 = P(A_i)P(A_j) = (1/2)^2$ when i and j are distinct. Thus the collection A_1, A_2, A_3 is pairwise independent. But $P(A_1A_2A_3) = 1/4 \neq P(A_1)P(A_2)P(A_3) = (1/2)^3 = 1/8$ so the collection is not mutually independent.

An example of this sort where S consists of three or fewer points is not possible (Exercise 6) so the above example is the simplest possible in that it consists of the least number of points.

SUGGESTED READINGS

Parzen, Chapter 2, Section 4.
Goldberg, Chapter 2, Sections 5, 7, 8.
Feller, Chapter 5, Sections 1–3.

EXERCISES

5.1 In Example 2, for what values of n and i are U_i and W independent?

5.2 In roulette (as in Exercise 3.2), which of the following pairs of events are independent?

(*a*) First dozen, second dozen
(*b*) Red, black
(*c*) Even, low
(*d*) Odd, low
(*e*) First dozen, odd

If 0 and 00 are eliminated, so that there are just the 36 equally likely outcomes $1, \cdots, 36$, which of the pairs (*a*) to (*e*) are independent?

5.3 In a certain voting district, 50% of the voters are Democrats, 40% are Republicans, and 10% are unaffiliated. A bond issue is favored by 80% of the Democrats, 20% of the Republicans, and 45% of the unaffiliated. A voter is selected at random and found to favor the bond issue. What is the probability that he is a Democrat, Republican, or unaffiliated, respectively?

5.4 Suppose $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$, and $P(A \cup B) = \frac{2}{3}$. Are A and B independent? If instead $P(A \cup B) = \frac{3}{4}$, are A and B independent?

5.5 What is the probability that South is dealt a hand at bridge which contains no spades? North picks up his hand and notes that he has six spades. What now is the probability, as far as North is concerned, that South has no spades? West, the player on North's right, opens the bidding with "one spade." Assume for simplicity that this means West has precisely five spades. What now is the probability, as far as North is concerned, that South has no spades? If West's bid of "one spade" means that West has at least five spades, what do you think you can say about the probability that South has no spades? Can you formally prove your assertion?

The famous probability theorist Emile Borel (1871–1956) has written (with A. Cheron) a book on probabilities at bridge, *Theorie Mathématique du Bridge*, 2nd ed., Gauthier-Villars, Paris, 1955.

5.6 Prove that if S has three points, a collection of three sets A_1, A_2, A_3 with $0 < P(A_i) < 1$ cannot be pairwise independent. Deduce that for an example of a collection which is pairwise independent but not mutually independent, S must have at least four points.

5.7 If A and B are independent, A' and B, A and B', and A' and B', are independent also.

5.8 Two true dice are rolled. Given that at least one of them shows a six, what is the probability that both show a six?

5.9 A random sample of size $k < n$ is drawn with replacement from $S = \{1, \cdots, n\}$. If $1 \leq j < k$, what is the probability that 1 will be drawn at least j times, given that it is drawn $j - 1$ times?

5.10 If C is an event such that $P(C) > 0$, prove the following.

PC1. $P(S \mid C) = 1$.

PC2. $P(E \mid C) \geq 0$ for all $E \subset S$.

PC3. $P(E + F \mid C) = P(E \mid C) + P(F \mid C)$ for disjoint sets E and F.

Consequently we have shown that the function P_C from 2^S into the real numbers, defined by $P_C(E) = P(E \mid C)$, is a probability measure whenever $P(C) > 0$. In particular, all the properties which can be deduced for any probability measure from the axioms, such as (*a*)–(*e*) preceding Definition 3.1, hold for P_C.

Notice that for any set E, this new probability measure takes any set E and multiplies the old probability $P(EC)$ of the piece EC in C by $1/P(C)$ and multiplies the old probability $P(EC')$ of the piece EC' outside C by 0.

5.11 Let $A_1, \cdots, A_n$ be any events n such that $P(A_1 \cdots A_{n-1}) \neq 0$. Prove that $P(A_1 \cdots A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2) \cdots P(A_n \mid A_1 \cdots A_{n-1})$.

5.12 *Five-card draw poker.* What is the probability that player 1 is dealt a flush (event A)? What is the probability that player 2 is dealt a flush (event B), given A? Is $P(B \mid A) < P(B)$?

5.13 A collection of subsets of a given set S can be mutually (and hence pairwise) independent with respect to one probability measure P_1 and not pairwise (and therefore not mutually) independent with respect to another probability measure P_2. Give an example of this.

5.14 In Example 2, suppose the ith urn contains a total of n_i balls. Suggest a procedure for selecting a ball at random, that is, so all balls have equal probability of being selected.

6. BAYES' RULE

In Example 5.2, we have $n + 1$ urns numbered from 0 to n. The ith urn contains i white balls and $n - i$ black balls. We draw a ball at random from the total of $n(n + 1)$ balls. If the ball selected is white (event W), we found the probability $P(U_i \mid W)$ that the ball came from the ith urn (event U_i) is $2i/n(n + 1)$.

Before we examined the ball, we knew that the probability it came from the ith urn was $1/(n + 1)$. When we found that the ball was white, we could make the sharper (except when $2i = n$) inference that it came from the ith urn with probability $2i/n(n + 1)$.

This example illustrates a general situation in statistical inference. We are given a sample space S and a probability measure on it. In the example, the sample space consists of the elementary outcomes "one of the $n(n + 1)$ balls is selected." The probability measure is given by the statement that each of the balls has the same probability of being chosen. The sample space is partitioned into subsets and we take a sample from one of these subsets. In our example, the partition is the division of balls into urns, that is, the subsets are the collection of events $U_1, \cdots, U_n$. These subsets of the partition are sometimes referred to as subpopulations.

When we have taken the sample from the unknown subpopulation, we then examine it and use this information to improve our knowledge of which subset the sample was taken from. In our example, examination of the sample showed that event W had occurred. We then computed $P(U_i \mid W)$.

There is a general formula for the situation we have described. It is called Bayes' rule and was published posthumously in 1763 by the Reverend Thomas Bayes.

Theorem 1. *Bayes' Rule.* Let $H_1, \cdots, H_k$ be a partition of the probability space (S, P) with $P(H_i) > 0$ for all i. Let E be any event such that $P(E) > 0$. Then

$$P(H_i \mid E) = \frac{P(H_i)P(E \mid H_i)}{\sum_{i=1}^{k} P(H_i)P(E \mid H_i)}$$

Proof. $P(H_i \mid E) = P(H_iE)/P(E)$. Now $P(H_iE) = P(E \mid H_i)P(H_i)$, which transforms the numerator into the desired form. To transform the denominator, we have $P(E) = P(\sum_{i=1}^{k} EH_i) = \sum_{i=1}^{k} P(EH_i)$, using the fact that the EH_i are disjoint. Since $P(H_i) > 0$ for all i, $P(EH_i) = P(E \mid H_i)P(H_i)$ for all i. Substituting, we have $P(E) = \sum_{i=1}^{k} P(H_i)P(E \mid H_i)$, as desired.

In some applications, the H_i are referred to as "causes." The theorem is sometimes called Bayes' rule for the probability of causes. In other applications, the H_i are sometimes referred to as "hypotheses." As we explained in Section 1.1, probability theory, which is descriptive, is the antithesis of determinism, which explains things by cause and effect. The use of the cause terminology often adds extraneous implications to Bayes' theorem. Avoid the cause terminology unless it is specifically indicated.

We call the $P(H_i)$ the **a priori** (or beforehand) **probabilities of the hypotheses** or causes. We call the $P(H_i \mid E)$ the **a posteriori** (or afterwards) **probabilities of the hypotheses** or causes, **given the event** E.

Example 1. The probability space is as in Example 5.2, with $n \geq 2$. A sample of two balls is drawn without replacement from one randomly selected urn. Suppose both balls are white (event W). What is the probability $P(U_i \mid W)$ that they come from urn i?

The role of E in Bayes' rule is played by W and the roles of the H_i are played, respectively, by the U_i. There are $b(n,2)$ ways to select two balls from n balls, and $b(i,2)$ ways to select two white balls from i white balls. We therefore have $P(W \mid U_i) = b(i,2)/b(n,2) = (i)_2/(n)_2$. Since $P(U_i) = 1/(n + 1)$, we have, using Bayes' rule and simplifying

$$P(U_i \mid W) = (i)_2 \Big/ \sum_{i=0}^{n} (i)_2$$

By using Exercise 4, this becomes $3(i)_2/(n + 1)_3$.

Example 2. A probability student who frequents the beach estimates at today's weather will be sunny (event S) with probability 0.9, and that will be cloudy (event C) with probability 0.1. He listens regularly to a :ather forecaster and assigns to his predictions one of three interpreta- ›ns: sunny, cloudy, and uncertain. His extensive records show the llowing conditional probabilities of the various forecasts given the rious actual days' weather.

Actual days' weather:	Interpretation of forecast: sunny F_s	cloudy F_c	uncertain F_u
Sunny S	0.8	0.1	0.1
Cloudy C	0.4	0.4	0.2

›tice that the rows must add to 1, according to the expression for $P(E)$ in proof of Theorem 1. If the forecast for today is interpreted as sunny, at is the a posteriori probability that it will be sunny, assuming that all probabilities given are accurate?

Jsing Bayes' rule,

$$P(S \mid F_s) = \frac{P(S)P(F_s \mid S)}{P(S)P(F_s \mid S) + P(C)P(F_s \mid C)} = 18/19 \doteq 0.95$$

SUGGESTED READINGS

rant, *Differential and Integral Calculus*, Volumes I and II, Interscience,)37. See Section 1.4.1 for a discussion of the evaluation of $\sum_{i=1}^{n} i^k$, a problem milar to 6.4.

dberg, Chapter 2, Section 6.

:en, Chapter 3, Section 4.

EXERCISES

I A blackjack expert finds that he is cheated in 10% of the games in which ·lays. Suppose he plays for 1 hour or until he loses 20 units, whichever rs sooner. If the dealer cheats, he is certain to lose 20 units in an hour. If lealer does not cheat, the probability is 0.25 that he will lose 20 units in an . Suppose that in a particular instance the expert loses 20 units in an hour. t is the probability that he was cheated?

! If $C_1, \cdots, C_n$ is a partition of S and $E \subset S$, prove that $EC_1, \cdots, EC_n$)artition of E.

; Assume for simplicity that one coin in a million is two-headed (event H_1) that all other coins are fair (event H_2). Suppose that a coin is selected at om and tossed 30 times in succession. If the result is heads each time (event ·hat is the probability that the coin is true?

*6.4 Prove that $\sum_{i=1}^{n} (i)_2 = (n+1)_3/3$.

6.5 Refer to Example 1. Let B be the event "two black balls are drawn" and let M ("mixed") be the event "one black ball and one white ball are drawn." Find $P(U_i \mid B)$ and $P(U_i \mid M)$.

6.6 Refer to Example 2. Let $P(S) = p$ instead of 0.9. For what values of p will we have $P(S \mid F_s) > p$? What is the greatest value possible for $P(S \mid F_s)$?

6.7 Prove that $\sum_{i=1}^{n} (i)_r = (n+1)_{r+1}/(r+1)$.

7. INDEPENDENT TRIALS

Suppose n experiments are conducted and their outcomes are listed as an ordered n-tuple. If the sample spaces for the individual experiments are $S_1, \cdots, S_n$, then the n-tuples which describe the outcome of all the experiments have the form $(s_1, \cdots, s_n)$. The ith member s_i of the n-tuple is an element of the ith sample space S_i. When the S_i are arbitrary sets we have the following.

Definition 1. The cartesian product (after Descartes) $S_1 \times \cdots \times S_n$ or $\prod_{i=1}^{n} S_i$ of the collection $S_1, \cdots, S_n$ of sets is the set of all ordered n-tuples $(s_1, \cdots, s_n)$ where s_i is a member of S_i. The ith member of an n-tuple is called the ith coordinate of the n-tuple.

If S is any sample space, $E = \{s\}$ is an elementary event, and h is a function on 2^S, we sometimes write $h(s)$ for $h(\{s\}) = h(E)$. If $s = (s_1, \cdots, s_n)$ is an n-tuple, it is customary to further abbreviate $h((s_1, \cdots, s_n))$ as $h(s_1, \cdots, s_n)$. We call h **a function of n variables.**

Example 1. *The cartesian product of two copies of the real numbers.* Imagine a flat infinite plane or sheet. Pick any point O and call it the **origin.** Draw a pair of perpendicular lines through O and imagine them prolonged indefinitely. Let each line represent a copy of the real numbers. Call one line the horizontal or X_1 axis, and the other line the vertical or X_2 axis. From any point P in the plane, draw perpendiculars to the axes. Suppose the perpendicular to the X_i axis intersects it at the point x_i. Associate the 2-tuple (x_1, x_2) with P. This is illustrated in Figure 1.

There is a biunique correspondence between these 2-tuples of real numbers and the points of the plane. Thus the plane is a pictorial representation of the cartesian product of two copies of the real numbers. This correspondence allows us to convert the "pictorial" assertions of plane (or two-dimensional) Euclidean geometry into algebraic assertions about 2-tuples, and vice versa. It is often much easier to discover or prove a theorem in one context rather than in the other. This use of the cartesian product was introduced by Descartes (1596–1650). It is the forerunner and prototype of the general cartesian product.

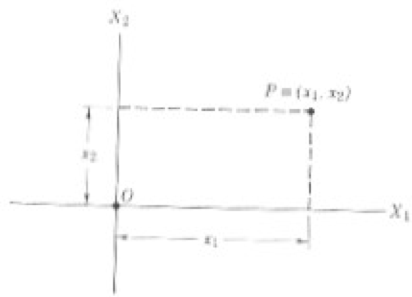
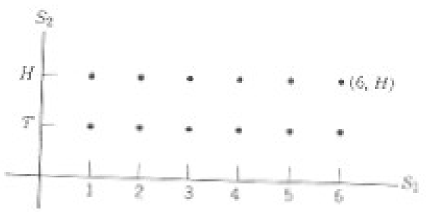


Figure 1. The cartesian product of two copies of the real numbers.

iolid, or three-dimensional, Euclidean geometry is studied similarly by ng a cartesian product $R^3 = R \times R \times R$ of three copies of the real nbers R. The power of the method becomes particularly apparent when study the geometry of four or more dimensions. Our pictorial talents there. But the algebraic machinery hums merrily on.

:xample 2. A die is rolled and then a coin is tossed. Here the cartesian duct $S_1 \times S_2$ consists of $6 \times 2 = 12$ points, as expected from Theorem 1. This is illustrated by Figure 2. The six die outcomes are indicated



'e 2. The cartesian product $S_1 \times S_2$, where S_1 is the sample space when a die lled and S_2 is the sample space when a coin is tossed.

ertical ticks numbered 1 to 6 on the horizontal or S_1 axis. The two omes of the coin toss are indicated by two horizontal ticks, T and H, he vertical or S_2 axis. The twelve 2-tuples of the cartesian product are :ated by dots in the plane.

discussing cartesian products, the following subsets will play a cularly important role.

:finition 2. Let $E_i \subset S_i$, $i = 1, \cdots, n$. The cartesian product $\cdots \times E_n$ or $\prod_{i=1}^n E_i$ is the set of all n-tuples $(s_1, \cdots, s_n)$ such that in E_i for all i.

Example 3. (*a*) In Example 1, let $E_1 = \{x_1 : x_1 \geq 0\}$, $E_2 = \{x_2 : x_2 \geq 0\}$. Then $E_1 \times E_2$ is the upper-right quadrant, including the non-negative axes.

(*b*) In Example 2, let $E_1 = S_1$ and $E_2 = T$. Then $E_1 \times E_2$ is the lower row of dots. If, instead, $E_1 = \{6\}$ and $E_2 = \{H\}$, then $E_1 \times E_2 = \{(6, H)\}$.

The subsets of Definition 2 ma·· be visualized as "generalized rectangles." If any E_i is empty, the product is empty. If we specify a subset $\prod_{i=1}^n E_i$ of a cartesian product by only de scribing m of the E_i, then it is understood that there is no restriction on the remaining E_i, that is, they are equal to the corresponding S_i. In the special case where at most one of the E_i is a proper subset of the corresponding S_i, that is, the product set is of the form $S_1 \times \cdots \times S_{i-1} \times E_i \times S_{i+1} \times \cdots \times S_n$, we abbreviate the product set by $E_i{}^*$. Using this abbreviation, we have for any cartesian product set $\prod_{i=1}^n E_i = \bigcap_{i=1}^n E_i{}^*$.

Of course, not all subsets of $\prod_{i=1}^n S_i$ are generally of the form $\prod_{i=1}^n E_i$. In Example 1, the interior and boundary of any (nontrivial) triangle is an example of a set that is not a product set. In Example 2, there are $2^{12} - 1 = 4095$ distinct nonempty subsets. There are only $(2^6 - 1)(2^2 - 1) = 63 \times 3 = 189$ distinct nonempty product sets.

The cartesian product describes for us the joint outcome of n experiments. We still have the problem of assigning a probability measure P to the cartesian product. If we are given probability measures P_i for the individual sample spaces S_i, we would expect to be able to use these P_i to determine P. This is, however, not possible in general.

There is an important special case in probability theory where P can be determined from the P_i. This is when the experiments are "independent." Consider the situation in Example 2, where a die is rolled and a coin is tossed. We might expect the outcome of each of these experiments to in no way affect the probabilities of the various outcomes of the other experiment. Thus we expect the cartesian product measure P to satisfy $P(E_1{}^*E_2{}^*) = P(E_1{}^*)P(E_2{}^*)$, the condition for independence. But $P(E_i{}^*) = P_i(E_i)$ and $E_1{}^*E_2{}^* = E_1 \times E_2$ so we have $P(E_1 \times E_2) = P_1(E_1)P_2(E_2)$. Thus, if we know P_1 and P_2, we can compute P for any product set. The probabilities of a joint event for the two independent experiments is the product of the probabilities of the individual events for each experiment.

These ideas extend to the general case of n experiments.

Definition 3. Let (S_i, P_i), $i = 1, \cdots, n$, be probability spaces representing n experiments or trials. The trials are independent if the probability measure P for the sample space $S = \prod_{i=1}^n S_i$ of joint outcomes of these experiments satisfies $P(\prod_{i=1}^n E_i) = \prod_{i=1}^n P_i(E_i)$ for every product set

$_{i=1}^{n} E_i$. To indicate that S has the product measure P defined above, and t P depends on the P_i, we may write $(S,P) = \prod_{i=1}^{n} (S_i,P_i)$. The product asure alone may be indicated by $P = \prod_{i=1}^{n} P_i$.

There is precisely one probability measure P which satisfies the con-ons of Definition 3. To see this, note that any set consisting of a single aple $\{(s_1, \cdots, s_n)\}$ can be written in the form $\{s_1\} \times \cdots \times \{s_n\}$ and refore is a product set. Since P is specified on product sets, it is specified sets consisting of one point. As we remarked after Definition 3.1, P is v determined on all subsets of $S_1 \times \cdots \times S_n$ because of $P3$.

n the particular case where the outcomes in each experiment are equally ly, we see that each set consisting of one n-tuple has the same proba-:y. (Our assumption at the beginning of Section 1.3 that each of the 36 iples is equally likely when two dice are rolled, is now seen to follow n assuming that each die is true and that the dice fall independently of another.)

simple but important special case of independent events has played a ral role in probability theory.

efinition 4. Let $A_1, \cdots, A_n$ be mutually independent events on (S,P) ı that $P(A_i) = p$, $P(A_i') = q = 1 - p$, for each i. Then $A_1, \cdots, A_n$ called Bernoulli trials with probability of success p, and probability ailure q.

ossing a coin n times, where heads represent success and tails represent re, are an example of Bernoulli trials. Suppose $T = \{0,1\}$ is any ability space of two points, and the probability measure Q on T is ied by $Q(0) = q$, $Q(1) = p$. Then the cartesian product $(S,P) =$ $_1 (T,Q)$ and the events A_i, where A_i is success on the ith trial, are oulli trials.

hen a true coin is tossed, the probabilist assumes that the trials are endent, that is, that they are Bernoulli trials with $p = 1/2$. Apparently ge students do not generally believe that the trials are independent. ads or tails guessing test of more than a thousand students by psychol- William Lepley was reported in "Looking into People," *Cosmo-in*, April, 1964, page 38. Lepley and his assistants called every throw s to see the reaction. Before the first toss, 77% guessed heads. The dropped somewhat to 66% before the second toss. When told that had come up for the second time, only 30% guessed heads before the toss. When the students were told that this result was also heads, gure rose to 42% guessing heads at the fourth toss! (A discussion of st and related misconceptions appears in "Probability," by Edward iorp, *Cosmopolitan*, September 1964, pp. 10–11.) Experiments in my lasses gave results almost identical to those of Lepley.

SUGGESTED READINGS

Bell, Chapter 3, for an account of the life and mathematical contributions of Descartes.

Abbott, *Flatland*. An entertaining story illustrating notions in two, three, and four dimensions.

EXERCISES

7.1 Clearly indicate by sketches the following subsets of Example 1.

(*a*) $1 \leq x_1 < 2$ and $3 < x_1 \leq 5$.
(*b*) $x_1 = 3$, $x_2 = -1$.
(*c*) $x_1 = 0$.
(*d*) $x_2 = 1$.
*(*e*) $x_1 x_2 \geq 1$.
*(*f*) $x_1^2 + x_2^2 < 1$.
(*g*) The exterior and boundary of a circle of radius 1 centered at the origin.

7.2 Clearly indicate by sketches the following subsets of Example 2, if possible.

(*a*) s_1 is even.
(*b*) $s_2 = H$.
(*c*) $s_1 = 1$ or 2, $s_2 = T$.
(*d*) E_1 is empty, $E_2 = S_2$.
(*e*) $\{(1,H),(2,T)\}$.

7.3 Which sets in the preceding exercises do not have the form $\prod_{i=1}^{n} E_i$?

7.4 A cubical die is generally labeled (spotted) so the sum of opposite faces equals 7. (There are still two nonequivalent ways to label a die, corresponding to the left and right-handed three dimensional coordinate systems.) One of the many forms of untrue dice (that is, crooked) which have been used in gambling casinos are Ace-six flats*. (The one on a die is often called an Ace.) The cube is shortened in the one-six direction. Suppose that a pair of Ace-six flats are rolled, where the probability of the various faces for each die is given by $P(\{1\}) = P(\{6\}) = 1/4$ and is $1/8$ for each of the other faces. Compute the probability of the various totals from 2 to 12. Compare the results for Ace-six flats with those for true dice listed in 1.3.

7.5 Refer to Example 3.1. Suppose the two coins each have probabilities a and b respectively of falling "heads." If the outcomes for the two coins are independent, prove that no allowable choices of a and b will justify Castor's probability measure.

*8. THE PROBABILITY OF THE UNION OF n EVENTS

It frequently happens that we are given the probabilities of n events $A_1, \cdots, A_n$ and wish to find $P(\bigcup_{i=1}^{n} A_i)$, whether or not the A_i are independent.

* The Silver Slipper casino on the famed Las Vegas strip was closed after Ace-six flats were found on its dice tables. (See *Los Angeles Herald Examiner*, "Crooked Dice Charge; Vegas Casino Closed," April 4, 1964, page 1.)

Theorem 1. Given the probability space (S,P), let $A_1, \cdots, A_n$ be any
llection of events. The probability that at least one of these events occurs

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \Sigma P(A_{i_1}) - \Sigma P(A_{i_1}A_{i_2}) + \cdots$$
$$+ (-1)^{(r-1)} \Sigma P(A_{i_1} \cdots A_{i_r}) + \cdots + (-1)^{(n-1)} \Sigma P(A_{i_1} \cdots A_{i_n})$$

ere each sum is taken over all $b(n,r)$ possible subsets of distinct indices
$< \cdots < i_r$. In particular the last sum reduces to $(-1)^{(n-1)}P(A_1 \cdots A_n)$.
roof. There are several well-known proofs. Two are given as Exercises
nd 2. A third can be based on the ideas of Exercise 4, which is a special
e of that proof.

xample 1. *Matching.* Suppose a permutation of the integers from 1
is chosen at random. What is the probability P_n that for at least one i
ith integer is in the ith position?
et A_i be the event "the ith integer is in the ith position." Then $P(A_i) =$
$-1)!/n! = 1/n$ and in general $P(A_{i_1} \cdots A_{i_k}) = 1/(n)_k$.
hus

$$_n \equiv P\left(\bigcup_{i=1}^{n} A_i\right) = \frac{b(n,1)}{n} - \frac{b(n,2)}{(n)_2} + \cdots$$
$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!}$$

e

$$1 - e^{-1} = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!} + \cdots$$

ave

$$p_n \doteq 1 - e^{-1} = 0.632 \cdots$$

act, $1 - e^{-1}$ lies between p_n and $p_n + (-1)^n/(n+1)!$ so we have an
r of less than $1/(n+1)!$ in the approximation.

ample 2. (*a*) A deck of n cards is well shuffled. The chance that
y card will be displaced from its original position is $1 - p_n \doteq 1/e$.
) Each of n guests bring a surprise package to a party. The packages
huffled and redistributed randomly. The probability that no one gets
his own package is $1 - p_n \doteq 1/e$.

ample 3. A true coin is tossed six times. What is the probability
the sequence $\cdots HTH \cdots$ appears?
t A_i, $i = 1, \cdots, 4$, be the event "$\cdots HTH \cdots$ appears, beginning in
." Then $P(A_i) = 1/8$. $P(A_iA_j) = 0$ if $j = i + 1$, $1/32$ if $j = i + 2$, and
$A_4) = 1/64$. The terms containing three or more A_i are 0. Thus the
d probability is $4/8 - 2/32 - 1/64 = 27/64$.

SUGGESTED READINGS

Feller, Chapter IV, and Parzen, II.6, give more extensive discussions and additional problems.

EXERCISES

8.1 Prove Theorem 1 by mathematical induction.

8.2 Prove Theorem 1 by the following informal counting argument.

(*a*) Consider those points of S which are in none of the A_i. These points make no contribution to any of the terms on either side of the equation.

(*b*) Consider those points which are contained in precisely one of the A_i. Each one contributes once to the left side of the equation. Note that the same is true of the right side.

(*c*) Consider those points which are contained in precisely k of the A_i. Each one contributes once to the left side of the equation and contributes $\sum_{i=1}^{k} (-1)^{i-1} b(k,i) = 1 - (1-1)^k =$ one time to the right side.

(*d*) We have partitioned S into "cases" and found that the equation is true for each case separately. It follows from $P3$ that the equation is true.

8.3 In a game of bridge, let A_i be the event that a randomly dealt hand is void (contains no cards) in the ith suit. What is the probability that this randomly dealt hand will be void in at least one suit?

8.4 If E is any subset of the finite set S, let $N(E)$ equal the number of points in E.

(*a*) If C_E is the characteristic function of E, then $N(E) = \sum_{s \in S} C_E(s)$.

(*b*) If $A = \bigcup_{i=1}^{n} A_i$, then $A' = \bigcap_{i=1}^{n} A_i'$ and $C_A = 1 - C_{A'} = 1 - \prod_{i=1}^{n} C_{A_i'}$ $= 1 - \prod_{i=1}^{n} (1 - C_{A_i}) = \sum_i C_{A_i} - \sum_{i_1,i_2} C_{A_{i_1}} C_{A_{i_2}} + \cdots$ where the sums are as in Theorem 1.

(*c*) $N(A) = \sum_{s \in S} C_A(s) = \sum_{s \in S} \sum_i C_{A_i}(s) - \sum_{s \in S} \sum_{i_1,i_2} C_{A_{i_1}A_{i_2}}(s) + \cdots =$ $\sum_i N(A_i) - \sum_{i_1,i_2} N(A_{i_1}A_{i_2}) + \cdots$, which is the same as the equation of Theorem 1 with P replaced by N.

(*d*) Dividing through by $N(S)$ yields Theorem 1 for the special case of the equiprobable measure.

CHAPTER 3

Random Variables on Discrete Sample Spaces

1. DISCRETE SAMPLE SPACES*

Now we extend probability theory to certain experiments with infinitely many outcomes. The reasons for doing this are both practical (Example 1) and conceptual (Example 2).

Example 1. *Radioactivity.* The atoms of certain of the chemical elements decay radioactively. This means that they "explode," emitting either a particle or a photon (packet of energy), and change into a different kind of atom. There is no way to predict the precise time that an atom of radioactive substance will decay. Experiments show that about half of the atoms of a substance will decay in a given time characteristic of the substance, which is known as the half-life. The probability that any given atom does not decay in one half-life is $\frac{1}{2}$, and in n half-lives it is $1/2^n$. If our experiment is to measure the time it takes for a specified atom to decay, the outcomes of the experiment may be thought of as the non-negative real numbers. Thus we have a sample space consisting of an infinite number of points.

Example 2. Suppose we toss a coin until a tail appears. Our sample space S of outcomes includes the positive integers $\{1, 2, \cdots\}$—the number of tosses required. Notice that the methods of Chapter 2 tell us that the probability of $H, H, \cdots, H, T$ with a tail on the nth toss, is $p^{n-1}q$, where p is the probability of a head, $0 < p < 1$, and $q = 1 - p$ is the probability of a tail. There is one more conceivable outcome, $H, H, \cdots, H, \cdots$ forever, which must be included in S. We label it ∞. What value should we give to $P(\{\infty\})$? In order for ∞ to occur, we must have $H, \cdots, H$ for each sequence of n trials. Therefore we expect $P(\{\infty\}) \leq P(H, \cdots, H) = p^n$ for all n. But this is only possible if $P(\{\infty\}) \leq 0$. Since probabilities ought to be non-negative, we get $P(\{\infty\}) = 0$.

* The reader will need elementary facts about convergence of series. In particular he should understand the geometric and exponential series and that the convergence and sum of a series of non-negative terms are unaffected by arbitrary grouping or rearrangement.

It seems natural to define $P(E)$ for any $E \subset S$ as the sum (finite or infinite) of the probabilities of the individual points in E. This yields $P(S) = \sum_{i=1}^{\infty} P(\{i\}) = q \sum_{i=1}^{\infty} p^{i-1} = q/(1 - p) = 1$, where we use the fact that the sum of the **geometric series** $1 + x + x^2 + \cdots$ is $1/(1 - x)$ when $|x| < 1$. (It further turns out that the sum is not defined otherwise.)

These ideas carry over to a certain kind of infinite sample space. A set S is called **countably infinite** if there is a biunique correspondence between S and the positive integers. In other words, the elements of the infinite set S can be completely labelled by using the positive integers. An infinite set which is not countably infinite is **uncountably infinite,** or **uncountable.** Countably infinite sets are the smallest infinite sets in the sense that every infinite set contains a (in fact, many) countably infinite subset. The idea of the proof is to label successive elements of the infinite set S with the successive positive integers. The process can be continued indefinitely, otherwise S is finite, contrary to assumption.

Definition 1. A discrete sample space $S = \{s_1, \cdots, s_n, \cdots\}$ is one which is either finite or countably infinite.

The set of rational numbers and the set of all integers are examples of countably infinite sets. The set of real numbers, or any nontrivial interval thereof, are examples of uncountably infinite sets. The sample space $S = \{t \geq 0\}$ of Example 1 is of this kind. It will be handled later, in Chapter 4.

Our ideas for finite sample spaces extend to discrete sample spaces much as in Example 2. Suppose $S = \{s_1, \cdots, s_n\}$. Let $P(\{s_i\}) = p_i \geq 0$ with $\Sigma\, p_i = 1$. Define $P(E) = \sum_{s \in E} P(\{s\})$ for any event E. We do not have to specify the order in which we combine the terms of the (possibly infinite) sum. This follows from the condition $p_i \geq 0$. It is a fact about infinite sums of positive terms $p_1 + \cdots + p_n + \cdots$ that they always sum up to the same value no matter in what order they are added (or, rearranged: when we arrange the terms of a sum in the order $p_1 + \cdots + p_n + \cdots$, it is understood that terms are successively added on in that order).

Note that the Axiom $P1$ of Section 2.3 is satisfied and that $P2$ follows at once from the definition of $P(E)$ and the fact that a finite or infinite sum of non-negative terms is non-negative. For two disjoint sets E and F, $P3$ asserts that

$$\sum_{s \in E} P(\{s\}) + \sum_{s \in F} P(\{s\}) = \sum_{s \in E+F} P(\{s\})$$

This is a consequence of the more general fact about series, where the probabilities are replaced by arbitrary non-negative numbers. We obtain $P2$ from this by mathematical induction. We used a stronger property than $P2$ when we defined $P(E)$ as a possibly infinite sum of probabilities of elementary events. When E is an infinite set composed of elementary events $E_1, \cdots, E_n, \cdots$, we let $P(E) = P(E_1) + \cdots + P(E_n) + \cdots$.

Definition 2. Suppose I is an index set (that is, set of labels) and $\{E_i\}_{i\in I}$ is a collection of subsets of a set S, labeled by the elements of I. (Usually, I will be a subset of the integers.) The union $\bigcup_{i\in I} E_i$ of the collection is the set of all $s \in S$ such that s is in at least one E_i. The intersection $\bigcap_{i\in I} E_i$ of the collection is the set of all $s \in S$ such that s is in every E_i. The collection is disjoint, or a subpartition, if $E_iE_j = \varnothing$ whenever $i \neq j$. A union of disjoint events is sometimes written $\sum_{i\in I} E_i$ or, when I is the positive integers, $E_1 + \cdots + E_n + \cdots$. The reader should verify the following.

(*a*) The following statements are equivalent for a collection $\{E_i\}_{i\in I}$. (1) The collection is pairwise disjoint. (2) $E_i \cap (\bigcup_{j\neq i} E_j) = \varnothing$ for every i. (The collection is "mutually disjoint.")

(*b*) de Morgan's laws and the distributive laws hold for arbitrary unions and intersections.

(*c*) The definitions in Section 2.3 of **partition** and **cover** extend to arbitrary collections of sets by replacing $\bigcup_{i=1}^{n} A_i$ by $\bigcup_{i\in I} A_i$ throughout.

The stronger property which we used to define $P(E)$ may now be phrased as follows. If $\{E_i\}$ is a finite or infinite collection of disjoint elementary events,

$$P(E_1 + \cdots + E_n + \cdots) = P(E_1) + \cdots + P(E_n) + \cdots$$

Using facts about series of positive terms, it can be proven that the P thus defined satisfies:

P4 *countable additivity*. For any countably infinite collection $\{E_i\}$ of disjoint events, $P(E_1 + \cdots + E_n + \cdots) = P(E_1) + \cdots + P(E_n) + \cdots$.

Definition 3. A probability measure for a discrete sample space S is a real-valued function P defined on 2^S which satisfies $P1$, $P2$, and $P4$.

The reader should verify the following.

(*a*) Axiom $P3$ follows from $P4$ by choosing $E_i = \varnothing$ from some point onward.

(*b*) Any function P defined on the elementary events is then determined on all subsets by applying $P4$. The determination does not depend on how points are labeled because the infinite sums are unaffected by arrangements. Thus it suffices to specify the probability measure on elementary events in defining a probability measure on a discrete sample space.

(*c*) Any function P which satisfies $P1$, $P2$, and $P4$ is of the type described following Definition 1.

(*d*) Since $P4$ implies $P3$, the axioms $P1$, $P2$, and $P4$ are stronger than $P1$ to $P3$. The only difference was that in the latter case, S had only a finite number of points. Therefore all our definitions and theorems about probability measures on finite sample spaces will also be true on the larger class of discrete sample spaces, provided that a finite number of points is not required.

In stating and proving theorems for discrete probability spaces, we will generally discuss only the infinite case. This is because we have either given the proof already in Chapter 2 or because the proof in the finite case is an obvious modification of the proof in the infinite case.

It is a convenience in proofs to note that we can assume without loss of generality that $S = \{1, 2, \cdots\}$. This is because a countably infinite set of outcomes can always be labeled by the positive integers, if desired. It also simplifies proofs to note that all elementary events having probability zero can be discarded from our sample space if desired (see Exercise 11). We shall therefore generally assume in proofs that $p_i > 0$ for all i.

(*e*) Definition 2.5.1, conditional probability, is unchanged for discrete probability spaces.

(*f*) Bayes' rule extends to the cases of finite and countably infinite partitions of discrete spaces (Exercise 6).

We make the obvious changes to extend Definition 2.5.2.

Definition 4. For any probability space (S,P), a collection of (two or more) events $\{A_i\}_{i\in I}$ is **mutually independent with respect to (S,P)** if $P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$ for every finite subset $i_1, \cdots, i_k$ of distinct indices.

In Example 2.4.2 we gave an experimental way to "realize" any finite probability space, that is, a real experiment for which the given (S,P) was the appropriate model. The same thing can be done for discrete sample spaces. The coin toss experiment of Example 1.2 does this when the probabilities for the nth outcome are given by $p^{n-1}q$, $0 < p < 1$, $q = 1 - p$. The general case is given in the next example.

***Example 3.** Suppose (S,P) is an infinite discrete probability space with probabilities $\{p_i\}$. We assume that all the $p_i > 0$. We make a composite experiment, consisting of a sequence of experiments, each with finitely many outcomes.

Choose a real number r such that $0 < r < 1$. Consider the finite probability space (S,P_1) with probabilities of the elementary outcomes given by $q_i = p_i$, $1 \le i < n_1$, $q_{n_1} = \sum_{i \ge n_1} p_i < r$. We determine n_1 by the restriction $q_{n_1} < r$. Realize this probability space by an experiment. If the experiment yields one of the first $n_1 - 1$ outcomes, we stop. If it yields the n_1 outcome, we consider the finite probability space (S,P_2) with probabilities of the elementary outcomes given by $q_i = p_i/q_{n_1}$ when $n_1 \le i < n_2$, and $q_{n_2} = \sum_{i \ge n_2} p_i/q_{n_1} < r$, which determines n_2. For later convenience, we have labeled the elementary outcomes for (S,P_2) by the integers from n_1 to n_2. We realize (S,P_2) by an experiment. If the outcome is other than the n_2 one, we stop. If it is the n_2 one, we consider a third finite (S,P_3) and construct a further experiment. The conditions are $q_i = p_i/q_{n_1}q_{n_2}$, $n_2 \le i < n_3$, and $q_{n_3} = \sum_{i \ge n_3} p_i/q_{n_1}q_{n_2} < r$, which determines n_2.

We assume that the successive experiments are independent. Then the probabilities $q_i = p_i$ for all i and the probability that we will have to conduct more than k experiments is less than r^k (Exercise 9). Thus the probability is one that only finitely many experiments are required to reach a decision. The smaller we are able to choose r, the more rapidly we are able to reach a decision. This is the only reason for introducing r to determine the n_i. If we do not care how quickly we reach a decision, the choice $n_i = i + 1$ will suffice (Exercise 10).

EXERCISES

1.1 Describe the probability measure in Example 2 for the cases $p = 0$ and $p = 1$.

1.2 If S is countably infinite, there is no equiprobable measure on S.

****1.3** A coin is tossed until a sequence of n consecutive heads (a **run** of **length** n) occurs. Describe a discrete sample space S of outcomes and define a suitable probability measure on S.

1.4 A coin has positive probabilities p, q, and r of falling heads, tails, or on edge, respectively. Suppose that the coin is tossed repeatedly and independently until either a head or a tail appears. Prove that $P(H) = p/(p + q)$, and $P(T) = q/(p + q)$. (See Example 1.3.1.)

1.5 A pair of true dice are rolled repeatedly. For each $t = 4, 5, 6, 8, 9, 10$ (**the field**), find the probability that t will come up before 7 comes up. What is the probability that 7 will come up before a specified t? What is the probability that neither 7 nor a specified t will come up?

1.6 Prove Bayes' rule (Theorem 2.6.1) for the cases of finite and countably infinite partitions of discrete spaces. The statement and proof for finite partitions is unchanged. For countably infinite partitions we have: **Bayes' rule.** Let $H_1, \cdots, H_n, \cdots$ be a partition of the discrete probability space (S,P) with $P(H_i) > 0$ for all i. Let E be any event. Then

$$P(H_i \mid E) = P(H_i)P(E \mid H_i) \Big/ \sum_{i=1}^{\infty} P(H_i)\, P(E \mid H_i).$$

1.7 A countably infinite collection of urns is numbered $1, 2, \cdots$. The nth urn contains 1 white ball and $n - 1$ black balls. An urn is selected in such a way that the probability of selecting the nth urn is $1/2^n$. (Suggest a procedure for doing this.) Then one ball is drawn from the urn. If the ball is white, find the probability that it was drawn form the nth urn.

****1.8** If S is any infinite set and P is a real-valued function on 2^S which satisfies P1 and P2, then $P(\{s\}) > 0$ for at most countably many $s \in S$.

***1.9** In Example 3, prove that $q_i = p_i$ for all the i and that the probability that more than k experiments will be required for a decision is less than r^k.

***1.10** If we let $n_i = i + 1$ in Example 3, prove that $q_i = p_i$ for all i and that the probability is one that a decision will be reached in a finite number of experiments.

1.11 Let S be a discrete sample space. Suppose that there are, among the outcomes of S an infinite number of elementary outcomes, each of which have probability zero.

(*a*) Using the fact that an infinite subset of a countably infinite subset is countably infinite, let $Z = \bigcup_{i=1}^{\infty} \{s_i\}$ where $s_i \in Z$ if and only if $P(\{s_i\}) = 0$.

(*b*) $P(Z) = 0$; and $P(E) = 0$ if and only if $E \subset Z$.

(*c*) For any $E \subset S$, $P(EZ) = 0$ and $P(EZ') = P(E)$.

Form the new probability space (T,Q), where $T = SZ'$ and $Q(E) = P(EZ')$ for $E \subset T$. This corresponds to discarding all sample points in which S have probability zero. The new probability space (T,Q) assigns the same probabilities as did the old probability space (S,P) to all elementary events with positive probability.

1.12 Two true dice are rolled repeatedly, until either a total of ten appears, in which case player 1 wins, or until a six appears, in which case player 2 wins. Describe a sample space for this game. Assign probabilities to the points of the sample space. Compute the probabilities that player 1 wins, that player 2 wins, and that neither wins. (Consider as "points" all finite sequences of rolls that lead to a resolution. Lump together all other (infinite) sequences of rolls as one elementary event.)

2. RANDOM VARIABLES

The result of an experiment is most often numerical. Even when it is not, the real numbers are so familiar and easy to work with that it is useful to convert the outcome of an experiment into numerical form.

Example 1. The vast area of measurement supplies many and varied illustrations.

(*a*) Many parameters of human populations are recorded as numbers: heights, weights, IQ, blood pressure, opinions. A finite sample space of the whole collection of individuals might be appropriate. Or the sample space might be some special subpopulation chosen either for convenience (for example, draftees—the army has certain standard records for them) or for special interest (for example, identical twins are often chosen to study the effects of differing environment and the same heredity on various parameters).

(*b*) The measurements of the physical sciences, such as of the speed of light, the half life of a type of atom, the rate of a chemical reaction, the mass of a planet, the strength of a magnetic field, the angular separation of a telescopic binary star—all generally result in numbers. It has been asserted that the only experimental data that are relevant to physical theories are pointer readings.

(*c*) The vast collections of statistics can be regarded as the results of experiments on suitable sample spaces. Examples are the voluminous data from business and economics (as one small case, the daily records of stock market transactions), and medical and actuarial statistics.

Example 2. In certain industrial and agricultural quality control situations, items are often classified simply as acceptable or defective. The experiment is examining a sample. The numerical result is the fraction of defectives observed in the sample.

Example 3. In a particularly simple example, n dice are rolled and we are interested in the (numerical) total that comes up.

In each situation where the result of the experiment is numerical and we can construct a suitable sample space S for the experiment, we can consider a function $X\colon S \to R$, where R is the real numbers, and the function is defined by letting $X(s)$ be the numerical result associated with the outcome s of the experiment. In Example 3, if the sample space were the set of ordered n-tuples $(i_1, \cdots, i_n)$, the function would be given by

$$X((i_1, \cdots, i_n)) = i_1 + \cdots + i_n.$$

Definition 1. If (S,P) is a discrete probability space, a random variable is a real-valued function X defined on S. The sum $X = \Sigma X_i$ of a finite collection $\{X_i\}$ of random variables on S is the random variable defined by $X(s) = \Sigma X_i(s)$ for each s in S. The product $Y = \Pi X_i$ of a finite collection of random variables on S is the random variable defined by $Y(s) = \Pi X_i(s)$ for each s in S. If c is a real number and X is a random variable, the random variable $Z = cX$ is defined by $Z(s) = cX(s)$ for each s in S. The real numbers are sometimes called scalars and the random variables cX are often called scalar multiples of X. A constant random variable C is one such that $C(s) = c$ for some real number c and all s in S.

We shall designate random variables by capital letters throughout. Random variables derive their usefulness from the fact that operations of addition, multiplication, and scalar multiplication closely resemble many of the properties of addition and of multiplication of real numbers.

The term "variable" in mathematics generally refers to a quantity which can assume various values. In the expression $X(s)$, where X is a function, both s and $X(s)$ are variables. The "argument" s of the function is called

the **independent variable** and the point $X(s)$ is called the **dependent variable**, for the obvious reasons. Thus, "random variable" is a misleading name for the function X. Unfortunately the name is established in the literature. A better term is "random function."

Remark. A subset N of a probability space (S,P) such than $P(N) = 0$ is called a **null set** or a set of **measure zero.** A probability space which describes a given situation may or may not have nonempty sets of measure zero. In Example 1.2, where we toss a coin until a tail appears, $\{\infty\}$ is such a set and it is the only such set, provided $0 < p < 1$. In Exercise 1.1, we saw that when $p = 0$, the null sets are precisely the subsets of $\{1\}'$, and when $p = 1$, the null sets are precisely the subsets of $\{\infty\}'$.

In discrete probability spaces, null sets can be eliminated simply by dropping from S all s in S such that $P(\{s\}) = 0$. However, it may be natural or convenient to allow such points. In both the theory and the practical applications, the addition or deletion of such points is immaterial.

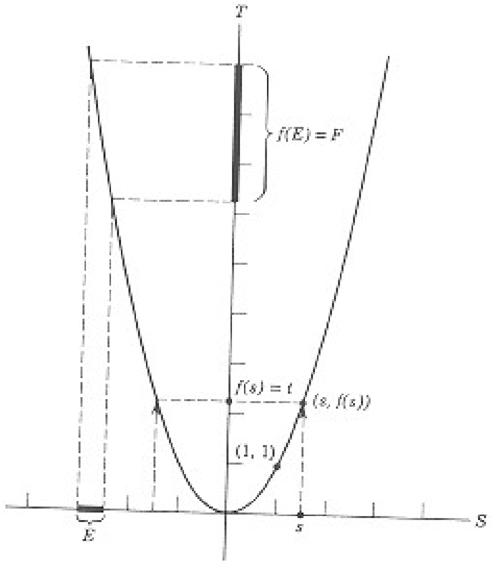
In particular, the values of a random variable may be arbitrarily specified on any nonempty null set. In a discrete space, $N = \{s\colon P(\{s\}) = 0\}$ is the largest null set. In defining a random variable, therefore, we need not specify it on N. If two random variables are equal except perhaps on N, we say they are **essentially equal.** If a random variable is a constant, except perhaps on N, we say it is **essentially constant.** Statements such as $X = Y$ or $X = C$ will always mean **essential equality** in what follows.

(*a*) The collection of real-valued functions on any set S is a vector space* under the usual operations of addition and scalar multiplication of functions: $(f + g)(s) = f(s) + g(s)$ and $(cf)(s) = cf(s)$, for all s in S and all real c. The vector space is furthermore an algebra with the usual operation of multiplication of functions: $(fg)(s) = f(s)g(s)$ for all s in S. A sign for the multiplication operation is omitted. It is indicated just as with real numbers by juxtaposition of f and g. The algebra has a unit, the constant function defined by $1(s) = 1$ for all s in S. The space of all random variables on a discrete probability space, with the operations of Definition 1, is a particular instance of this.

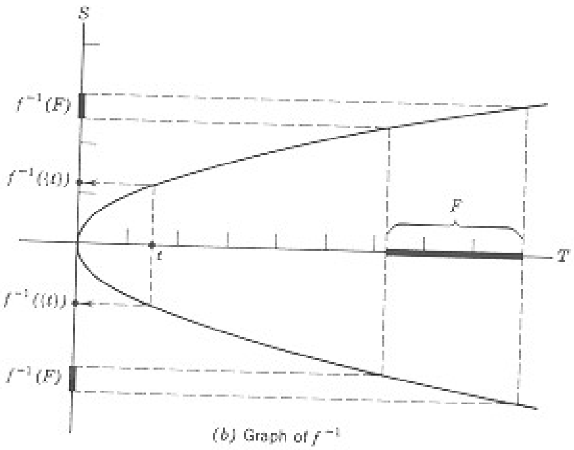
In order to examine the connection between a random variable X on S and the probability measure P on S, we need the concept of the inverse of a function.

Definition 2. Suppose the function $f\colon D \subset S \to R \subset T$ is given. The inverse of f is the function $f^{-1}\colon 2^T \to 2^S$ defined by $f^{-1}(F) = \{x \in S\colon f(x) \in F\}$. The set $f(E)$ is the image of E under f and $f^{-1}(F)$ is the counter-image, or inverse image, of F under f^{-1}.

* Knowledge of vector spaces in general, and of this observation in particular, are not needed at any point of this text. This is usually true of asterisked portions of the text.



(a) Graph of f



(b) Graph of f^{-1}

Figure 1. The graphs of f and f^{-1}, where $f(s) = s^2$. The image of each point of S is one point of T. The function maps two distinct points onto each positive number in T. The inverse function maps each positive number in T into a set of two points in S. Both f and f^{-1} map zero onto itself.

The reader should observe that, regardless of the domain or range of f, the domain of f^{-1} is always all of 2^T. Also, $f^{-1}(\varnothing) = \varnothing$ and $f^{-1}(T) = D$ always.

Notice that whereas f is a function which maps points of S into points of T, f^{-1} maps subsets of T into subsets of S. If f is 1–1, then whenever $F = \{t\} \subset R$ consists of one point, the reader should verify that $f^{-1}(\{t\})$ is also a set consisting of one point. Furthermore, if $t_1 \neq t_2$, and both t_1 and t_2 are in R, $f^{-1}(\{t_1\})$ and $f^{-1}(\{t_2\})$ are disjoint, that is, they consist of distinct one point subsets. In this case, we can define a function $g: Y \to X$ with domain R by choosing, for each t in R, the one $g(t)$ such that $\{g(t)\} = f^{-1}(\{t\})$. This function is 1–1. It is customary to call the 1–1 function g by the name f^{-1}. Thus the term f^{-1} is used to denote either of two separate and distinct functions, one from Y to X and one from 2^Y to 2^X. We can distinguish between them by determining what the domain and range spaces are for the f^{-1} currently being discussed.

Similarly, we associate with any function f the set function $g: 2^S \to 2^T$ defined by $g(E) = \{f(s): s \in E\}$. We shall also denote this g by f.

Example 4. Let S and T be the real numbers and let $f: S \to T$ be defined for all s in S by $f(s) = s^2$. The **graph of any function** $f: S \to T$ is defined as the set of ordered pairs $(s, f(s))$ of the cartesian product $S \times T$. The **graph of** f^{-1} is then defined as the set of ordered pairs $(f(s), s)$ of $T \times S$. Figure 1 illustrates the graphs of f and f^{-1}. In particular, it is easy to "see" various facts about f and f^{-1} from the sketches.

In experiments where one is interested in numerical results, it is the various probabilities of these results that are of interest, rather than an underlying probability space (S,P). If S is finite, any random variable X on S has a finite set of real numbers as its range, $\{x_1, \cdots, x_m\}$. Associated with each x_i is a probability q_i that x_i will be the value of X. The q_i can be calculated if S,P, and X are given. But if P, for example, were unknown, we could estimate the q_i by repeating the experiment and using relative frequencies. We could even do this if both S and P were unknown, provided merely that we assume there is an underlying probability space. (There may be many probability spaces which are equally appropriate for a given experiment, as we pointed out in Example 2.3.2.) This is the actual situation in numerous applications in probability and statistics: interest is in a certain set of numerical values and we either do not know or do not care to know an appropriate (S,P).

If S is countably infinite, the range of any X is either a finite or a countably infinite set. The discussion is the same except that an experiment can now only estimate accurately a finite number of the q_i.

Definition 3. Let (S,P) be a discrete probability space and let X be a random variable on S. The frequency or density function f_X for X is defined for each x_i in $X(S)$ by

$$f_X(x_i) = P(X^{-1}(\{x_i\})) \equiv P(X^{-1}(x_i)) \equiv P(X(s) = x_i) \equiv P(X = x_i)$$

The first expression involving P is the notationally precise way to define f_X. The others are successively less precise synonyms that are often used. The distribution function F_X for X is defined for all real x by $F_X(x) = \sum_{x_i \leq x} f_X(x_i) = P(\{s\colon X(s) \leq x\}) \equiv P(X(s) \leq x) \equiv P(X \leq x)$. Again, the first expression involving P is the most precise and the others are successively less precise synonums.

Example 5. Let X be the number of heads when a coin is tossed n times. We have $P(X = k) = b(n,k)p^kq^{n-k} = f_X(k)$. Any random variable with this density function is said to be **binomially (n,p)-distributed.**

A density function f for a random variable on a discrete probability space satisfies the following:

d1· $f(x_i)$ is defined and $f(x_i) \geq 0$ for each x_i where $\{x_i\}$ is a finite or countably infinite set of real numbers. (If $f = f_X$, this set is the range of X.)

d2· $\Sigma f(x_i) = 1$.

Conversely, any such density function is the density function for some random variable. To see this, suppose f satisfies $d1$ and $d2$. Let $S = \{x_1, x_2, \cdots\}$ and let $P(x_i) = f(x_i)$. Then $P(S) = \Sigma f(x_i) = 1$ by $d2$, and $d1$ insures that the usual definition of $P(E)$ as $\sum_{x_i \in E} f(x_i)$ gives $P(E) \geq 0$ for all E. The definition of $P(E)$ automatically gives $P4$. Hence P is a probability measure for the discrete space S. Now let I be the random variable on (S,P) defined by $I(x_i) = x_i$ for all i. Then

$$f_I(x_i) = P(\{x\colon I(x) = x_i\}) = P(\{x_i\}) = f(x_i)$$

Therefore f is the density function for the random variable I. Since every density function is therefore the density function for one random variable (and, in general for many), we are led to the following.

Definition 4. Any function f which satisfies $d1$ and $d2$ is a discrete density function or frequency function. Any function F which can be written as $F(x) = \sum_{x_i \leq x} f(x_i)$, where f is a discrete density function, is a discrete distribution function.

The graph of a density function is a scattered collection of points in the plane. These points are often joined to the horizontal axis by vertical lines, creating a histogram effect, as in Figure 2.

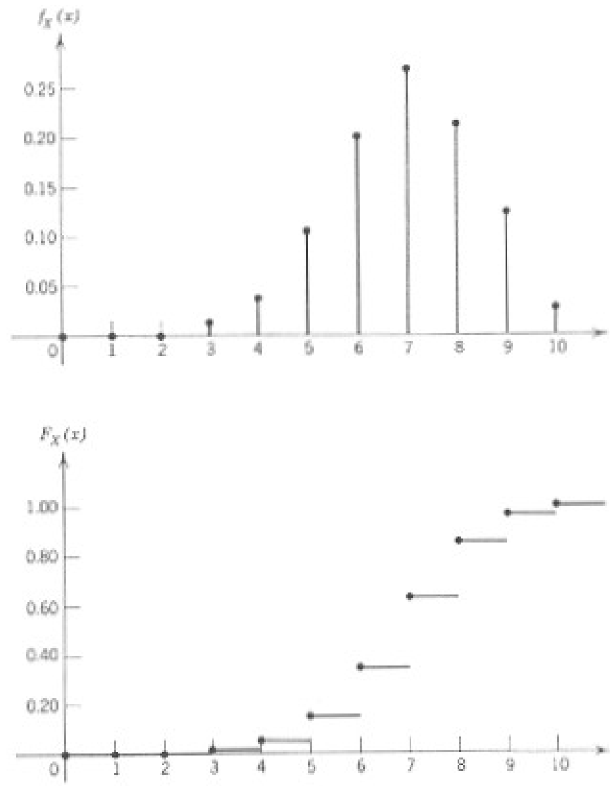


Figure 2. Graphs of the density and distribution functions for a binomially distributed random variable with $n = 10$, $p = 0.7$. The graph of F_X is zero for all $x < 0$ and is one for all $x \geq 10$. The discontinuities or "jumps" at $x = 0, 1, \cdots, 10$ are of magnitude $f_X(x)$. The point x belongs for these values to the right horizontal segment but is missing from the left ones.

Of particular importance are collections of random variables which are independent in the sense that knowledge of the values of some members of the collection imply nothing about the probabilities of the various values of other members.

Definition 5. A finite collection $\{X_i\}$ of random variables on a discrete sample space (S,P) is pairwise independent if for every collection $\{F_i\}$ of subsets of the real numbers, $\{X_i^{-1}(F_i)\}$ is a pairwise independent collection of sets. Otherwise the collection is pairwise dependent. The random variables are mutually independent (or, simply independent) if $\{X_i^{-1}(F_i)\}$ is a mutually independent collection for every $\{F_i\}$. Otherwise the random variables are dependent. An arbitrary collection of random variables is pairwise independent or mutually independent, respectively, if every finite subcollection has the property.

We shall be primarily interested in mutually independent random variables.

Theorem 1. Let $X_1, \cdots, X_n$ be a finite collection of random variables on the discrete sample space (S,P). Then $\{X_i\}$ is mutually independent if and only if for each set of real numbers $\{x_i\}$, with x_i in the range of X_i for each i, $P(X_1 = x_1, \cdots, X_n = x_n) = \Pi P(X_i = x_i)$.

Proof. Suppose the X_i are independent. Then

$$\begin{aligned} P(X_1 = x_1, \cdots, X_n = x_n) &= P(\cap X_i^{-1}(x_i)) \\ &= \Pi P(X_i^{-1}(x_i)) = \Pi P(X_i = x_i). \end{aligned}$$

* We sketch the proof of the other part of the theorem. Suppose $F_1, \cdots, F_n$ is a collection of subsets of real numbers. We wish to show that $\{X_i^{-1}(F_i)\}$ is an independent collection of subsets of S. Since S is discrete, the range $X_i(S)$ is finite or countably infinite so $G_i = F_i \cap X_i(S)$ is also finite or countably infinite. We write $G_i = \{x_{i,1}, x_{i,2}, \cdots\}$. Also $X_i^{-1}(F_i) = X_i^{-1}(G_i)$. Now each G_i is a disjoint union of one point subsets so for each i we have $X_i^{-1}(G_i) = \sum_k X_i^{-1}(x_{i,k})$, using the first equation in Exercise 15*a* and noting that the union is of disjoint sets. Thus

$$P(X_i^{-1}(F_i)) = P\Big(\sum_k X_i^{-1}(x_{i,k})\Big) = \sum_k P(X_i^{-1}(x_{i,k})) = \sum_k p_{i,k}$$

where we let

$$p_{i,k} = P(X_i^{-1}(x_{i,k}))$$

Now

$$\begin{aligned} P(\cap X_i^{-1}(F_i)) &= P(\cap X_i^{-1}\{x_{i,1}, x_{i,2}, \cdots\}) \\ &= P\Big(\sum_{k_1, \ldots, k_n} \bigcap_i (X_i = x_{i,k_i})\Big) = \sum_{k_1, \ldots, k_n} P(\cap (X_i = x_{i,k_i}) \end{aligned}$$

and by hypothesis, each term in the latter can be replaced by a product, giving $\sum_{k_1, \ldots, k_n} \prod_i p_{i,k_i}$. These are precisely the same terms that arise from multiplying out $\prod_i \sum_k p_{i,k}$. Since all the terms are non-negative, they can be rearranged without affecting the expressions. Thus $P(_i \cap X_i^{-1}(F_i)) = \Pi P(X_i^{-1}(F_i))$. The argument is the same when the complete set of indices $i = 1, \cdots, n$ is replaced by any finite subset. Thus the $X_i^{-1}(F_i)$ are mutually independent and since the F_i are arbitrary, the X_i are independent.

The reason for giving Definition 5, rather than the simpler equivalent property established in Theorem 1, as the definition for independent random variables, is that the former is the form which is suitable for later generalization beyond discrete probability spaces. We shall generally use the easier property of Theorem 1 to check for independence of discrete random variables.

Example 6. In n tosses of a coin, with $0 < p < 1$, let X_i be the number of heads at the ith trial. Then $X_i = 0$ or 1 depending on whether there was a tail or a head on the ith trial. The total number of heads is $S_n = X_1 + \cdots + X_n$. It is readily seen from Theorem 1 that the X_i are independent random variables. However S_n and X_i, for some fixed i, are not independent. For example, $P(S_n = 0, X_i = 1) = 0 \neq P(S_n = 0)P(X_i = 1) = q^n p$.

The particular random variable that we may choose as the result of an experiment is somewhat arbitrary. For example, we could measure height in centimeters, inches, or various other units. A change of units replaces a random variable by a scalar multiple of itself. More complex changes occur when we transform raw data by numerical functions. Instead of values $x_1, \cdots, x_n$, we might list instead $\log x_1, \cdots, \log x_n$ or perhaps $\sqrt{x_1}, \cdots, \sqrt{x_n}$.

We formalize this process of making new random variables from old ones by real-valued functions defined on the real numbers.

Definition 6. The composition $h = g \circ f$ of the functions $f: S \to T$ and $g: T \to U$ is the function $h: S \to U$, with domain $f^{-1}(g^{-1}(U))$ and range $g(f(S))$, defined for each s in its domain by $h(s) = g(f(s))$.

In particular, if f is a real-valued function with domain D_f a subset of the real numbers, and X is a random variable on S with $X(S) \subset D_f$, then $f \circ X$ is also a random variable on S. Whenever we speak of $f \circ X$ in the sequel, it is understood that $X(S) \subset D_f$. The next theorem shows us that if a collection of random variables is independent, the new collection obtained by applying such functions is still independent, as we should expect intuitively.

Theorem 2. Suppose $X_1, \cdots, X_n$ are independent random variables on (S,P), that $f_1, \cdots, f_n$ are real-valued functions with domains in the real numbers, and that $X_i(S) \subset D_{f_i}$ for each i. Then the $f_i \circ X_i$ are also independent on (S,P).

Proof. Let $F_1, \cdots, F_n$ be subsets of the real numbers. Then

$$(f_i \circ X_i)^{-1}(F_i) = X_i^{-1}(f_i^{-1}(F_i)) = X_i^{-1}(G_i)$$

and the latter are independent sets because the X_i are independent random variables. The conclusion follows. (The first equality follows from Exercise 19.)

EXERCISES

2.1 Let X be nonempty and let $Y = \{1\}$. If f is any function from X to Y (each choice of $D \subset X$ gives a different f), find f^{-1}. Are there functions $g: 2^Y \to 2^X$ which are not an f^{-1} for any f, even though g has domain 2^Y?

2.2 Let $D = X = \{1,2\}$ and $Y = \{3,4\}$. Define f by $f(1) = f(2) = 3$. Evaluate f^{-1} for each point (that is, set) in the domain of f^{-1}.

2.3 A true coin is tossed 100 times. Let X be the number of heads which appear. Find the density and distribution functions, f_X and F_X.

2.4 A bridge hand is dealt from a complete well shuffled deck. Let X be the number of honor cards (A,K,Q,J,10) in the hand. Find the density and distribution functions f_X and F_X.

2.5 Mathematical models for the stock market often assume that, if $X_1, \cdots, X_n, \cdots$ are successive daily closing prices, the differences $\Delta X_i = X_{i+1} - X_i$, $i = 1, \cdots, n-1$, are mutually independent random variables. In order to test this assumption people have (inconclusively, as it happens) examined the random variables $Y_i = \operatorname{sgn}(\Delta X_i)$, where sgn is the function from the real numbers to the real numbers defined by $\operatorname{sgn} t = 1$ if $t > 0$, $\operatorname{sgn} t = -1$ if $t < 0$, and $\operatorname{sgn} 0 = 0$.

(*a*) If the Y_i fail to be mutually independent, what can we conclude about the mutual independence of the ΔX_i? Why?

(*b*) If the Y_i are mutually independent, must the ΔX_i be mutually independent? Why?

***2.6** If $(S,P) = (S_1,P_1) \times \cdots \times (S_n,P_n)$ is a cartesian product of finite spaces and $\{X_i\}$ is a collection of random variables such that X_i depends only on the ith coordinate, that is, $X_i(s_1, \cdots, s_n) = f_i(s_i)$, then the X_i are independent on (S,P).

2.7 Let an urn contain n balls numbered from 1 to n. Suppose a sample of size k, $2 \le k \le n$, is drawn without replacement. Let X_i be the number of the ball drawn on the ith trial. Are the X_i even pairwise independent?

2.8 (*a*) An arbitrary collection of constant random variables is independent.

(*b*) If a collection of random variables is independent, it remains independent when any collection of constant random variables is included with it.

(*c*) A subcollection of an independent collection of random variables is independent.

(*d*) Suppose $\{E_i\}$ are subsets of the discrete probability space (S,P) and $\{C_{E_i}\}$ are the corresponding characteristic functions. Then the E_i are pairwise or mutually independent sets if and only if the C_{E_i} are respectively pairwise or mutually independent.

(*e*) Pairwise independence and mutual independence are equivalent for two random variables but mutual independence is a stronger condition on collections of three or more, in general.

2.9 If the f_i of Theorem 2 are 1 to 1, show that the converse holds, that is, independence of the $f_i \circ X_i$ implies independence of the X_i. Thus, in testing for independence, we can apply 1-1 functions to the ranges of the X_i to transform them into a more convenient form.

2.10 Let $X_1, \cdots, X_m$ be independent random variables on a discrete probability space (S,P) such that $P(s) > 0$ for all s. If the range of X_k contains at least n_k points, then S contains at least $\prod n_k$ points.

In particular, if S has n points [and $P(s) > 0$ for all s], then a collection of independent random variables (none of which are constant) has at most $\log n/\log 2$ members.

2.11 Give simple examples of the following.

(*a*) $\{E_i\}$ is a cover of D_f and $\{f(E_i)\}$ is not a cover of R_f.

(*b*) $\{E_i\}$ is a partition of S and $\{f(E_i)\}$ is not a partition of $f(S)$.

(*c*) $\{E_i\}$ is a partition of $D_f \ne S$ and $\{f(E_i)\}$ is a partition of the range space T of f.

(*d*) For given integers $m \ge n > 0$, a function f such that D_f has m points and R_f has n points. A function f such that D_f is infinite and R_f has n points.

2.12 Given functions f, g, and h, for which $g \circ f$ and $h \circ g$ are defined, prove the associative law $h \circ (g \circ f) = (h \circ g) \circ f$.

****2.13** Some authors mean by a discrete probability distribution one where the density function f is positive only for "isolated" values. A set of values is isolated if for every real number t, whether or not it belongs to the set, there is an interval surrounding t of the form $\{x: |x - t| < d\}$ for some d and which contains no member of the set other than t. In other words, there is no real number at which the values accumulate or cluster.

(*a*) Prove that a set of values which is isolated is countable. Therefore this definition of discrete is more specialized than our definition.

(*b*) The set of values $1, \frac{1}{2}, \frac{1}{3}, \cdots$ is not isolated.

(*c*) Let $f(1/n) = 2^{-n}$, $n = 1, 2, \cdots$. This density function is discrete in our sense but not in the one above. Hence the one above is a proper subcase of our definition.

(*d*) For another example, note that no rational number is an isolated point of the set of rationals.

(*e*) For this definition show that the following set of axioms completely describe the distribution functions, that is, show that any function satisfying the axioms is a distribution function for this definition and, conversely, any distribution function for this definition satisfies these axioms.

D1 F is a real-valued function defined on the entire real line.

D2 F is monotone nondecreasing, that is, $x_1 < x_2$ implies $F(x_1) \le F(x_2)$.

D3 $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

D4 F is continuous except at a set of isolated points, where it is discontinuous. At its points of continuity, F is constant, that is, for each point of continuity c there is a $d > 0$ such that F has the same value on $\{x: |x - c| < d\}$. At its points of discontinuity, F is right continuous, that is, if c is a point of discontinuity, $\lim_{x \to c^+} F(x) = F(c)$, where the limit is from the right.

This characterization of distribution functions is one reason for introducing the definition above rather than the one we have used. We shall see in Section 4.4 a reason for choosing ours instead of the one above.

2.14 Transformations of the dummy variable for a finite sum, of the special form $j = k + b$ with b a nonzero integer and j and k dummy variables, were discussed in Section 1.4. Show that if $\sum_{k=m}^{n} a_k$ is a finite sum and f is a 1-1 function on the set S of integers k such that $m \le k \le n$, then $\sum_{k=m}^{n} a_k = \sum_{k=m}^{n} a_{f(k)} = \sum_{j \in f(S)} a_j$. Thus any 1–1 transformation f of the dummy variable leaves the finite sum unchanged.

2.15 (*a*) Prove that f^{-1} and the Boolean operations commute, that is,

$$\bigcup_{i \in I} f^{-1}(A_i) = f^{-1}\left(\bigcup_{i \in I} A_i\right), \bigcap_{i \in I} f^{-1}(A_i) = f^{-1}\left(\bigcap_{i \in I} A_i\right)$$

and $f^{-1}(A') = f^{-1}(A)'$.

(*b*) $f^{-1}(A - B) = f^{-1}(A) - f^{-1}(B)$.

(*c*) $f^{-1}(F) = \varnothing$ if and only if $F \cap f(S) = \varnothing$. $f(E) = \varnothing$ if and only if $E \cap f^{-1}(T) = \varnothing$. The function f is from S to T.

(*d*) If $\{F_i\}$ is a partition or a cover of the range R of f, or even of T, then $\{f^{-1}(F_i)\}$ is a partition or a cover, respectively of the domain of f.

(*e*) $f^{-1}(f(E)) \supset D(f) \cap E$. Give an example where the inclusion is proper. $f(f^{-1}(E)) = R(f) \cap E$.

2.16 Sketch the trigonometric functions sin (not sin x, a number), cos, tan, and their inverses. Discuss them (domain, range, whether the inverse is many-valued, etc.).

2.17 (*a*) For what value or values is a given (n,p)-binomial density function a maximum?

(*b*) Show that the binomial density function always decreases as k increases above the value (s) for which the maximum is attained, and also that the density decreases as k decreases below these values.

2.18 Give an example where $X_1, \cdots, X_n$ are mutually independent on (S,P_1) and not even pairwise independent on (S,P_2).

2.19 Refer to Definition 6. Prove that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$, and thus $(g \circ f)^{-1}(G) = f^{-1}(g^{-1}(G))$.

3. THE EXPECTATION OF A RANDOM VARIABLE

Associated with some (but not all) random variables on discrete spaces is a very important quantity $E(X)$ known as the expected value of the random variable X.

Definition 1. The expectation (or expected value) $E(X)$ of the random variable X on the discrete sample space (S,P) is $\sum_{s \in S} X(s)P(s)$, provided the sum is absolutely convergent. If the sum is not absolutely convergent, $E(X)$ is not defined.

To get a feeling for the significance of $E(X)$, consider a game of Nevada roulette. There are many bets available. For instance a player may bet on odd, even, red, black, high, or low. If any of the 18 numbers in the corresponding group come up, the bettor receives an amount equal to the amount he bet. If any of the 20 other numbers comes up, he loses. A

player may also bet on one of the dozens (1–12, 13–24, 25–36). If the dozen he bets on comes up, he receives twice his original bet. Otherwise he loses his bet. The dozens pay 2 to 1, or 3 to 1. A bet on a single number pays 35 to 1. There are several other bets.

In general, both for roulette and for the other casino gambling games, there are a variety of bets, each with a list of payoffs for the various possible outcomes. These bets are random variables. For Nevada roulette, the sample space is $\{1, \cdots, 36, 0, 00\}$. If the wheel is fair, P is the equiprobable measure. Then the result of a bet on a single number i is the random variable X defined by $X(i) = 35$, $X(j) = -1, j \ne i$. $E(X) = -\frac{1}{19} \doteq -5.26\%$.

What does $E(X)$ tell us about the game? Suppose we believe that for a large number n of trials the relative frequencies of the various outcomes will each be quite close to the probabilities of these outcomes (in this case 1/38 for each number). If one unit is bet each time, then the total loss or gain G_n for the player after n trials satisfies $G_n = 35W_n - L_n$, where W_n is the number of wins in n trials and L_n is the number of losses. The average win per trial is $G_n/n = 35W_n/n - L_n/n \doteq 35(\frac{1}{38}) - (\frac{37}{38})$. We have replaced the actual fraction of wins, W_n/n, and the actual fraction of losses L_n/n, by the good approximations 1/38 and 37/38, respectively. Thus $E(X)$ would appear to be the limiting value of the average gain per trial as the number n of trials increases. We shall later see that under suitable hypotheses this is a provable statement.

In the roulette example above, $E(X) \doteq G_n/n = (x_1 + \cdots + x_n)/n$, where x_i is the value of the random variable that was observed on the ith trial. This expression $(x_1 + \cdots + x_n)/n$ is known as the **arithmetic mean** or **average** of the numbers $x_1, \cdots, x_n$. We will see that under suitable hypotheses on the random variable X, the observed average will tend to $E(X)$ as n increase. Various forms of this fact are known as the "law of averages" to the layman (they are generally badly misunderstood by him), and as the various laws of large numbers to the probabilist. These laws are of central importance for statistics.

If $E(X)$ is defined, then $\Sigma\, X(s)P(s)$ is absolutely convergent so the terms can be grouped or rearranged in any order. If we group those terms having the same value of $X(s)$ and factor out this value of $X(s)$ from each group, we have $E(X) = \Sigma\, x_i P(X = x_i) = \Sigma\, x_i f_X(x_i)$. In particular, if X is a random variable with finite range (for instance, if S is finite), then this expression has only finitely many terms and is trivially absolutely convergent. Absolute convergence of this expression and the defining expression for $E(X)$ are equivalent. In fact, this expression may be used instead in Definition 1 and the result is an alternate but equivalent definition of $E(X)$ (see Exercise 1). Thus $E(X)$ is defined whenever X has finite range.

Example 1. Let S be the positive integers and let $P(n) = 2^{-n}$ for each n in S. Suppose X is defined by $X(n) = 2^n$ for each n. Then the expression for $E(X)$ is $1 + 1 + \cdots$ which does not converge so $E(X)$ is not defined.

An absolutely convergent series is convergent but a convergent series need not be absolutely convergent. Thus the expression $\Sigma x_i f_X(x_i)$ for $E(X)$ may converge but not absolutely.

Example 2. Choose (S,P) as in Example 1. Define X by $X(n) = 2^n(-1)^n/n$. Then $\Sigma X(n)P(n) = \Sigma(-1)^n/n$, a convergent series. But $\Sigma 1/n$ is not convergent so the series is not absolutely convergent.

Series which converge, but not absolutely, are called **conditionally convergent.** Such series always have an infinite number of positive terms and an infinite number of negative terms. Furthermore, the sum of the positive terms alone and of the negative terms alone diverge to $+\infty$ and $-\infty$, respectively.

If a conditionally convergent series is rearranged, the series may converge to another value or it may diverge to either $+\infty$ or $-\infty$. In fact, given any real number and any conditionally convergent series, the series may be suitably rearranged in many ways to converge to the given number. The series may also be rearranged to diverge to either $+\infty$ or $-\infty$ or to diverge without tending to either $+\infty$ or $-\infty$, that is, "oscillate." The fact that the sum of a conditionally convergent series always depends on the order of summation of the terms explains why $E(X)$ is not defined when the series $\Sigma X(s)P(s)$ is conditionally convergent: $E(X)$ should not depend on how we happen to list the outcomes in our description of the sample space.

Example 3. The series $\Sigma(-1)^n/n$ illustrates the behavior of a conditionally convergent series. It can be shown that the series as written converges to $-\log_e 2$. To see intuitively how it can be arranged to converge to any value, note that the positive terms are $\frac{1}{2}, \frac{1}{4}, \cdots$ and the negative terms are $-1, -\frac{1}{3}, \cdots$. The important facts are that there is an infinite supply of each of them and that they tend to zero. Suppose we wish the rearranged series to converge to some fixed real number r. Use up positive terms (in order) until the partial sum exceeds r. Now start using up negative terms in order until the partial sum is less than r. Then use more positive terms until the partial sum again exceeds r, and so forth. Use the positive and the negative terms in order, respectively. Of course, they are being used at different rates, in general, but they are each used once and only once.

For example, if we wish to rearrange $\Sigma(-1)^n/n$ so that it converges to 0, we would have at the end of each of the various stages described above:

$\frac{1}{2}$, $\quad \frac{1}{2} - 1$,
$\frac{1}{2} - 1 + \frac{1}{4} + \frac{1}{6} + \frac{1}{8}\ (= \frac{1}{24})$,
$\frac{1}{2} - 1 + \frac{1}{4} + \frac{1}{6} + \frac{1}{8} - \frac{1}{3}\ (= -\frac{7}{24})$, and so forth.

In general, the partial sums until the next stage is completed are always no farther from r than the last term of the previous stage. Since the terms tend to zero (a consequence of the convergence of the series), the partial sums tend to r.

It is not difficult to rearrange $\Sigma(-1)^n/n$ to diverge to $+\infty$, to $-\infty$, or to oscillate (Exercise 2).

These ideas can be converted into a formal proof of the correctness of our description of the behavior of a conditionally convergent series under rearrangement.

Example 4. If X is binomially (n,p)-distributed, $E(X) = np$. We have

$$E(X) = \sum_{k=0}^{n} kb(n,k)p^k q^{n-k}$$

and

$$kb(n,k) = nb(n-1, k-1), \quad k = 1, \cdots, n$$

Thus

$$E(X) = np\sum_{k=1}^{n} b(n-1, k-1)p^{k-1}q^{(n-1)-(k-1)} = np$$

The last sum is 1 because it is the sum of the terms of an $(n-1, p)$-binomially distributed random variable.

The most important properties of expectation are given in the next theorem.

Theorem 1. If X and Y are random variables on (S,P) whose expectations are defined, then for all real numbers a and b we have $E(aX) = aE(X)$, $E(X + Y) = E(X) + E(Y)$, and therefore (linearity) $E(aX + bY) = aE(X) + bE(Y)$. It follows by induction that for random variables $X_1, \cdots, X_n$ whose expectations are defined, and real numbers a_i, $E(\Sigma a_i X_i) = \Sigma a_i E(X_i)$.

Proof. $aE(X) = a\Sigma X(s)P(s) = \Sigma(aX(s))P(s) = \Sigma(aX)(s)P(s) = E(aX)$ where the parentheses strongly suggest the reasons for each step. The last two sums are absolutely convergent because absolute convergence of series is preserved when the terms are multiplied through by a scalar.

$$\begin{aligned} E(X) + E(Y) = \Sigma X(s)P(s) + \Sigma Y(s)P(s) &= \Sigma[X(s) + Y(s)]P(s) \\ &= \Sigma(X + Y)(s)P(s) = E(X + Y) \end{aligned}$$

The second equality and the absolute convergence of the third expression follow from the absolute convergence of the two sums in the second expression.

The rest of the theorem is proven readily.

Theorem 1 is perhaps the most important "easy" theorem in all of probability theory. Many otherwise complicated arguments become simple when it is used. Refer, for example, to Wilson, Appendix B.

Example 5. The computation in Example 4 is now trivial. Let $X = X_1 + \cdots + X_n$, where $P(X_i = 1)$ is p, $P(X_i = 0)$ is q, for each i. Then $E(X_i) = 0 \cdot q + 1 \cdot p = p$ and $E(X) = E(\Sigma X_i) = \Sigma E(X_i) = np$.

When $E(g(X))$ is defined, there is a simple and useful relation between $E(g(X))$ and the density function g_X of X (whether or not $E(X)$ is defined). It may happen that one of the numbers $E(X)$ or $E(g(X))$ is defined but the other is not (Exercise 3*c* and *d*).

Theorem 2. Let X be a random variable on the discrete sample space (S,P) and let g be a real-valued function such that $X(S) \subset D_g$. If $E(g(X))$ is defined, then

$$E(g(X)) = \Sigma g(X(s))P(s) = \Sigma g(x_i)g_X(x_i)$$

Proof. The first equality follows at once by substituting the random variable $g \circ X$ into the definition of expectation and using the definition of composition of functions. The second equality follows from grouping the terms corresponding to a given x_i. The factor $g(x_i)$ is common to each group. Each such group then yields the corresponding term in the last sum. The first sum can be so grouped without affecting its value because it is absolutely convergent.

Example 6. Let X be as in Example 4. Compute $E(X^2)$. Using Theorem 2, we have

$$E(X) = \sum_{k=0}^{n} k^2 b(n,k)p^k q^{n-k} = \sum_{k=1}^{n} k(k-1)b(n,k)p^k q^{n-k} + \sum_{k=0}^{n} kb(n,k)p^k q^{n-k} = n(n-1)p^2 \sum_{k=2}^{n} b(n-2, k-2)p^{k-2}q^{n-k} + np = (np)^2 + npq$$

If instead we use the approach of Example 5, we have

$$X^2 = (X_1 + \cdots + X_n)^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j$$

Now $X_i^2 = X_i$ so $E(X_i^2) = E(X_i) = p$ for each i. Since $X_i X_j$ is one when both are one and is zero otherwise, $E(X_i X_j) = p^2$, $i \neq j$. Thus

$$E(X^2) = \sum_{i=1}^{n} p + \sum_{i \neq j} p^2 = np + n(n-1)p^2 = (np)^2 + npq$$

as before.

The ease or difficulty of one method of calculation compared with another in mathematics varies with the problem and the individual.

EXERCISES

3.1 The definition of $E(X)$ for discrete (S,P) is commonly given using $\Sigma x_i f_X(x_i)$ in place of $\Sigma_{s \in S} X(s)P(s)$ in Definition 1.

(*a*) Show that if either sum converges absolutely, then the other does.

(*b*) Conclude that the definitions are equivalent, that is, the X for which $E(X)$ is defined are the same, and the values assigned to $E(X)$ are also the same.

3.2 Describe rearrangements of $\sum(-1)^n/n$ which diverge to $+\infty$, to $-\infty$, and which oscillate.

3.3 Give examples of the following and verify that they have the stipulated properties.

(*a*) A random variable X such that $E(X)$ is defined for (S,P_1) and is not defined for (S,P_2).

(*b*) $E(X) = 0$, X is not constant, and $P(s) > 0$ for all s.

(*c*) An f and an X, neither of which is constant, such that $E(X)$ is defined and $E(f(X))$ is not defined.

(*d*) An f and an X, neither of which is constant, such that $E(f(X))$ is defined and $E(X)$ is not defined.

3.4 If $\{A_i\}_{i \in I}$ and $\{B_j\}_{j \in J}$ are partitions of S, then $\{A_i B_j\}$ is a partition of S. **If I and J are countably infinite then there are countably many $A_i B_j$.

3.5 If X is a random variable, $P(\{s\}) > 0$ for each $s \in S$, and $E(|X|) = 0$, then $X = 0$.

3.6 A box contains two good flashlight batteries and two bad flashlight batteries which have been inadvertently and thoroughly mixed. Assume that the flashlight works if and only if two good batteries are inserted. Pairs of batteries are selected and tested until the pair is found that works. Let X be the number of tests which are made.

(*a*) Find f_X, F_X, and $E(X)$ when a different pair is tried each time.

(*b*) Find f_X, F_X, and $E(X)$ when the pair to be tested is selected "at random" each time, that is, there is no record of pairs already tested, and any pair has the same chance of being selected as any other for each and every test. Is it worth keeping a record?

(*c*) Generalize parts (*a*) and (*b*) to $n \geq 2$ good batteries and $n \geq 1$ bad batteries.

3.7 A set of real numbers is **bounded** if there is a (non-negative) number M such that $|t| \leq M$ for all t in F. If no such constant exists, F is **unbounded**.

(*a*) Show that a finite set is bounded.

(*b*) Give an example of an unbounded set.

(*c*) Prove that $E(X)$ is defined if the range of X is bounded. From (*a*) and (*c*) preceding, it follows then in particular that $E(X)$ is always defined for random variables with finite range, as we had observed prior to Example 1. Thus, this exercise generalizes that result.

3.8 If $E_1, \cdots, E_n$ are subsets of S, C_{E_i} are the characteristic functions of the E_i, and a_i are real numbers, then $E(\Sigma a_i C_{E_i}) = \Sigma a_i P(E_i)$.

3.9 If X is a real-valued function defined on a set S, let X^+ be the function defined by $X^+(s) = X(s)$ if $X(s) > 0$ and $X^+(s) = 0$ if $X(s) \leq 0$. Similarly, X^- is defined by $X^-(s) = X(s)$ if $X(s) < 0$ and $X^-(s) = 0$ if $X(s) \geq 0$.

(*a*) $X = X^+ + X^-$.

(*b*) $|X| = X^+ - X^-$.

(*c*) The following statements are equivalent for random variables:

(i) $E(X)$ is defined.

(ii) $E(X^+)$ and $E(X^-)$ are both defined.

(iii) $E(|X|)$ is defined.

3.10 If $\{A_i\}_{i\in I}$ is independent, so is $\{A_i, A_i'\}_{i\in I}$.

3.11 It is common to use the alternate definition of expectation given in Exercise 1. The proof in Theorem 1 that $E(X + Y) = E(X) + E(Y)$ becomes more difficult. Give a proof based on this alternate definition.

3.12 Let $X_1, \cdots, X_n$ be independent random variables on a discrete (S,P) such that $E(X_i)$ is defined for each i. Prove that $E(\Pi X_i) = \Pi E(X_i)$.

3.13 In European roulette, there are 36 numbers and only one zero, that is, there are 37 equally likely alternatives which are labeled $0, 1, \cdots, 36$. When the zero turns up, all stakes on the "even" chances (black, red, odd, even, etc.) are put in "prison." The player is now offered two alternatives: (1) He may settle at once for the return of half of his bet, or (2) the bets remain on the table and on the next spin of the wheel, if they "win," they are returned to the player without any additional payment. If zero comes up again, the player is again offered the two alternatives. Otherwise the bets are lost. What is the expectation on the even-chance bets?

Single number bets are paid 35 to 1 if they come up. Otherwise these bets are lost. What is the expectation of a single number bet?

3.14 Consider an infinite sequence of Bernoulli trials with probability p of success, $0 < p < 1$. The **first run** is the largest set of consecutive trials, beginning with the first trial, which have the same outcome as the first trial.

(*a*) Find the expected value of the length L of the first run. What happens when p is near zero or one? *For which p is $E(L)$ least? (Differentiation is helpful here.)

(*b*) The **second run** is the largest set of consecutive trials which have the same outcome, beginning with the trial that first differs from the initial trial. Explain intuitively why the expected length of the second run will generally be less than that of the first run.

(*c*) What is the expected length $E(L_2)$ of the second run?

**(*d*) Consider a doubly infinite sequence of Bernoulli trials (that is, indexed from $-\infty$ to ∞). What is the distribution of the length L of the run of which an arbitrarily selected trial, such as the zeroth, is a member? What is the expected value of L? What relation do you expect between this and the $E(L)$ and $E(L_2)$ computed previously?

3.15 In both the home and the casino game of craps two dice are rolled. If a total of 2, 3, or 12 appears on the first roll the bet **pass** loses. If 7 or 11 appear on the first roll, pass wins. If 4, 5, 6, 8, 9, or 10 appear on the first roll, the total which appears is called the **point**. The dice are then rolled until either the point or seven appears. If the point appears first, pass wins. If seven appears first, pass loses. The other principal bet is **don't pass**. In the home game this loses when pass wins and wins when pass loses. In the casino game there is one exception. If a twelve is rolled on the first trial, pass loses but don't pass neither wins or loses.

Find the expectation of the pass and don't pass bets, both for the home game and for the casino game. Compare Exercise 1.5.

3.16 Ace-six flats are dice that are thinner in the Ace-six direction so that these sides will show more often. Suppose $p_i = \frac{1}{6} + 2c$ when $i = 1$ or 6 and that $p_i = 1/6 - c$ when $i = 2, 3, 4$, or 5, for each of a pair of dice. Program a computer to solve Exercise 15 above for these "crooked" dice, for several values of c. Graph the results.

4. MOMENTS, VARIANCE, AND STANDARD DEVIATION

The expectation has long been known to physics (long before probability theory, for example, to Archimedes, 287–212 B.C.) as the center of gravity. Consider a unit mass distributed on the countable subset of the real line $\{x_1, x_2, \cdots\}$. Let p_i be the amount of mass at x_i. Then the moment of the mass p_i at x_i about a point c is defined as $p_i(x_i - c)$. Think of the real line as a horizontal rigid weightless "wire," fixed at c so that it can rotate about a horizontal axis perpendicular to it and through c, to which the masses p_i are affixed. Then the moment of p_i is (proportional to) the tendency of p_i to rotate the mass about the axis through c. The sum of the moments $M_c = \Sigma p_i(x_i - c)$ measures the tendency of the entire rigid mass distribution to rotate about c. If we think of ourselves as sitting far down the negative Y-axis, a net positive moment means the mass distribution will rotate clockwise and a net negative moment means the mass distribution will rotate counterclockwise. If c has the property that the net moment is zero, then c is called the **center of gravity** of the mass distribution.

Since the p_i are non-negative and $\Sigma p_i = 1$, $p_i = f(x_i)$ defines a density function f. If X is a random variable with range $\{x_1, x_2, \cdots\}$ and this density function, that is, $f_X = f$, then

$$M_c = \Sigma p_i(x_i - c) = \Sigma (x_i - c) f_X(x_i) = E(X - c) = E(X) - c$$

where c in the next to last expression stands for the constant random variable with range c, rather than the number c.

We have not yet defined $X - c$, where X is a real-valued function and c is a number. For convenience we now make the definition that $X - c$ is the real-valued function with the same domain as X and having the value $(X - c)(s) = X(s) - c$ for each s in the domain of X.

In particular, $M_c = E(X)$ when $c = 0$. For this reason $E(X - c)$ is called the (first) moment of X about c and $E(X)$ is the (first) moment of X about zero.

By analogy, we can define second and higher moments of random variables about points.

Definition 1. Let X be a random variable on a discrete sample space (S,P). The nth moment of X about the point c is $E[(X - c)^n]$, $n = 0, 1, 2, \cdots$, provided this quantity is defined. The rth absolute moment of X about the point c is $E(|X - c|^r)$, $r \geq 0$, provided this quantity is defined.

Similarly, if f is a density function, the nth moment of f about c is defined by $\Sigma (x_i - c)^n f(x_i)$ if this sum converges absolutely and the rth absolute moment of f about c is $\Sigma |x_i - c|^r f(x_i)$ if the sum converges.

We have defined the nth moments only for the non-negative integers because these are of primary interest and fractional powers of negative quantities lead to complications.

The second moment of a random variable also has an important physical equivalent. Consider a body moving at constant velocity (perhaps 0), that is, at a constant speed in a constant direction. The body resists any alteration of its speed or direction. This resistance is called inertia, and is used to define the mass of the object. Now consider a body rotating at a constant angular speed about a fixed axis. Similarly, the body resists any change in angular speed or in the direction of the axis of rotation. This resistance is an inertia for angular motion. It is proportional to the **moment of inertia** of the mass with respect to the prescribed axis, defined as $\Sigma p_i(x_i - c)^2$. In terms of random variables this is $E((X - c)^2)$, the second moment of X about c. This quantity is of particular importance in probability theory, as well as in physics.

Definition 2. The variance $V(X)$ of the random variable X is $E([X - E(X)]^2)$, the second moment of X about its mean, if this expectation is defined. The standard deviation $s(X)$ of X is then $V(X)^{1/2}$. Similarly, the variance of a random variable having the density function f is $V(f) = \Sigma [x_i - E(f)]^2 f(x_i)$ if the sum converges. We may speak loosely of $V(f)$ as the variance of the density function f. The standard deviation is then $s(f) = \sqrt{V(f)}$.

(*a*) $V(X) = E([X^2 - 2XE(X) + E(X)^2]) = E(X^2) - E(X)^2$.

(*b*) For a binomially distributed random variable, $V(X) = npq$ and $s(X) = \sqrt{npq}$.

(*c*) The zeroth moment of a random variable is 1.

(*d*) If m is a non-negative integer, the mth moment about a point exists if and only if the mth absolute moment about the point exists. Thus if we prove that an mth absolute moment exists, then we automatically know that the mth moment exists.

Theorem 1. If $X(S)$ is a bounded set of real numbers (that is for some M, $|X(s)| \le M$ for all s in S), then the absolute moments exist for all $r \ge 0$, consequently (see (*d*) above) the ordinary moments exist for all non-negative integers.

Proof. If c is a real number, $|X(s) - c| \le |X(s)| + |c| \le M + |c|$ for all s in S. Hence $\Sigma |X(s) - c|^r P(s) \le \Sigma (M + |c|)^r P(s) = (M + |c|)^r < \infty$ so the rth absolute moment is defined.

Of course the rth absolute moment of a given random variable about a point need not be defined when $r > 0$. The following example shows that this can happen for any $r > 0$.

Example 1. Let $S = \{1, 2, \cdots\}$, $P(n) = 2^{-n}$, $X(n) = 2^{n/r}$. Then the rth absolute moment of X about zero is not defined:

$$E(|X|^r) = \Sigma |X(n)|^r P(n) = \Sigma 2^n 2^{-n} = \infty$$

Note that the tth absolute moment is defined for all t such that $0 \le t < r$. This is a special instance of the situation described in the next theorem.

Theorem 2. If $E(|X - c|^r)$ is defined, then $E(|X - c|^t)$ is defined for all t such that $0 \le t \le r$ and consequently $E(|X - c|^k)$ is defined for all integers k such that $0 \le k \le r$.

Proof. If $|X(s) - c| \le 1$, then $|X(s) - c|^t \le 1$. If $|X(s) - c| > 1$, then $|X(s) - c|^t < |X(s) - c|^r$. Thus $|X(s) - c|^t \le 1 + |X(s) - c|^r$ for all s in S. Multiplying this inequality through by $P(s)$ and summing over all s in S gives $E(|X - c|^t) \le 1 + E(|X - c|^r)$. Since the latter sum converges (absolutely), so does the former, and the convergence is absolute.

Additional important properties of moments follow.

Theorem 3. $E(|X + Y|^r) \le E(|X|^r) + E(|Y|^r)$ if $0 \le r \le 1$ and $E(|X + Y|^r) \le 2^{r-1}(E(|X|^r) + E(|Y|^r))$ if $r > 1$. Consequently, if the rth absolute moments or the kth ordinary moments of X and Y exist about a point, so does the corresponding moment of $X + Y$.

If the rth absolute moment or the kth ordinary moment of X exists about one point it exists about every point. Thus, in discussing the existence of moments, we need not refer to points.

Proof. Let a and b be real numbers. For $0 \le r \le 1$, $|a + b|^r \le |a|^r + |b|^r$ and for $r > 1$, $|a + b|^r \le 2^{r-1}(|a|^r + |b|^r)$ (Exercise 1). Letting $a = X(s)$, $b = Y(s)$, multiplying each side of the inequalities by $P(s)$, and summing over all s in S, yields the inequalities to be established.

To prove the last assertion, suppose $E(|X - c|^r)$ exists. If d is any real number, $X - d = (X - c) + (c - d)$. Letting $Y(s) = c - d$ for all s and noting that the rth absolute moment of a constant always exists, it follows from the part of the theorem that has been proven that $E(|X - d|^r)$ exists.

Definition 3. The covariance* $\mathrm{cov}(X, Y)$ of two random variables X and Y, such that $E(X)$ and $E(Y)$ are defined, is $E([X - E(X)][Y - E(Y)])$, or $E(XY) - E(X)E(Y)$.

If $E(X)$ and $E(Y)$ exist, then so does the expectation of each factor, but the expectation of the product of the factors, that is, the covariance, may not exist (Exercise 12).

* The notion of covariance was introduced by Galton in 1885.

If X and Y are independent, $\operatorname{cov}(X, Y) = 0$. However, $\operatorname{cov}(X, Y) = 0$ is possible even though X and Y are not independent. Suppose for example that f_X is an **even function**, that is, $f_X(t) = f_X(-t)$ for all real t, and that $Y = X^{2n}$ for some integer $n > 0$. Then $E(XY) = \sum x_i^{2n+1} f_X(x_i) = 0$ and $E(X) = \sum x_i f_X(x_i) = 0$, so $\operatorname{cov}(X, Y) = 0$. But X and Y are not independent, provided that we exclude the trivial case where Y is constant. Intuitively this is obvious. For details, see Exercise 6.

To have $\operatorname{cov}(X, Y) = 0$ with X and Y dependent is not surprising when we consider that $\operatorname{cov}(X, Y) = 0$ imposes only one condition on X and Y. Independence, however, generally imposes many conditions on X and Y, corresponding to the various possible pairs of inverse images $X^{-1}(E)$ and $Y^{-1}(F)$ of sets of real numbers E and F that are to be tested for set independence.

Theorem 4. If $X = X_1 + \cdots + X_n$, where $E(X_i)$ is defined for each i, then $V(X) = \sum V(X_i) + \sum_{i \neq j} \operatorname{cov}(X_i, X_j)$. Consequently, if the X_i are pairwise independent, $V(X) = \sum V(X_i)$.

Proof.

$$E(X) = \sum E(X_i) \text{ so } V(X) = E(X - E(X))^2 = E(\sum [X_i - E(X_i)])^2.$$

Expanding the last expression, using the fact that the expectation of the sum of random variables is the sum of the expectations, and collecting terms, yields the result. The last assertion follows from the fact that the covariance of a pair of independent random variables is zero.

As we have seen, the covariance is a crude indicator of dependence between random variables. Roughly speaking, if the covariance is "large," dependence is "considerable." But as we have seen, small covariance (zero, for instance) does not mean independence. In talking of covariance, it is common to standardize the covariance so it can be compared in different situations.

Definition 4. If X and Y are random variables such that $E(X)$ and $E(Y)$ are defined, and $s(X)$ and $s(Y)$ are not zero (true if and only if neither X nor Y is constant), the correlation coefficient† is $k(X, Y) = \operatorname{cov}(X, Y)/s(X)s(Y)$.

The correlation coefficient may be rewritten

$$k(X, Y) = E\left(\left\{\frac{X - E(X)}{s(X)}\right\}\left\{\frac{Y - E(Y)}{s(Y)}\right\}\right)$$

where the quantities in curly brackets are the standardized random variables X^* and Y^* associated with X and Y, respectively (Exercise 5).

† The correlation coefficient was introduced by Karl Pearson in 1895.

Since the standardized random variables are not affected by a change of scale in X or in Y, the correlation coefficient is likewise unaffected.

Theorem 5. $|k(X, Y)| \leq 1$.

Proof. $0 \leq E((X^* \pm Y^*)^2) = E((X^*)^2) \pm 2E(X^*Y^*) + E((Y^*)^2) = 2(1 \pm \operatorname{cov}(X^*, Y^*)) = 2(1 \pm k(X, Y))$. Therefore $\pm k(X, Y) \geq -1$, which is equivalent to $|k(X, Y)| \leq 1$.

If $k(X, Y) = 1$, we see from the preceding that $V(X^* - Y^*) = 0$, that is, $X^* - Y^* = C$, where C is a constant random variable (Exercise 4). This can be rewritten as $Y = (s(Y)/s(X))X + D$, where D is another constant random variable. Thus Y is simply X under a change of scale. If the number pairs $\{(X(s), Y(s)) : s \in S\}$ are plotted in the XY-plane, the locus is a subset of a straight line with the positive slope $s(Y)/s(X)$. The relation between X and Y is said to be perfect linear positive. Conversely, if $Y = aX + b$, $a > 0$, it is readily seen that $k(X, Y) = 1$.

If $k(X, Y) = -1$, $X^* + Y^* = C$ and $Y = -(s(Y)/s(X))X + D$. Again Y is simply X under a change of scale. The relation between X and Y is called perfect linear negative. Conversely, if $Y = aX + b$, $a < 0$, it is readily seen that $k(X, Y) = -1$.

EXERCISES

4.1 Derive the inequalities used in the proof of Theorem 3.

4.2 Prove that $E((X - c)^2)$ is smaller when $c = E(X)$ than when c has any other value.

4.3 Give an example of a nonconstant random variable whose odd moments are all zero about the point zero.

4.4 If $E(|X - c|^r) = 0$ for one $r > 0$, then $X(s) = c$ for all s and

$$E(|X - c|^r) = 0$$

for all $r > 0$.

4.5 $V(aX) = |a|^2 V(X)$ and $V(X - c) = V(X)$. If X is not constant, let $X^* = (X - E(X))/s(X)$. Then $E(X^*) = 0$ and $s(X^*) = 1$. The random variable X^* is called the **standardized random variable** associated with X. The standardized random variable is unaffected by a linear change of scale, that is, $(aX + b)^* = X^*$, provided $a > 0$. If a < 0, $(aX + b)^* = -X^*$. If $U = aX + b$, $V = cY + d$, and $ac \neq 0$, then $k(U, V) = \operatorname{sgn}(ac)k(X, Y)$ where $\operatorname{sgn} t = 1$, $t > 0$, and $\operatorname{sgn} t = -1$, $t < 0$.

4.6 Verify the assertion following Definition 3 that, under the given hypotheses, X and X^{2n} are not independent. If $n = 0$, are they independent? If $f_X(0) = 1$, are they independent?

4.7 *The first absolute moment.*

(*a*) Graph the functions defined by $f(x) = |x|$, $|x|/2$, $|x|/3$, $|x - 1|$, $|x - 1|/2$, $|x - 1|/3$. At what points are these functions continuous? Differentiable?

*(*b*) Graph the function defined by $f(x) = |x|/2 + |x - 1|/3 + |x - 2|/6$. At what points is f continuous? Differentiable?

*(*c*) Where is the function in (*b*) a minimum?

(*d*) If X is a random variable with finite range $x_1 < \cdots < x_n$, let $g(c) = E(|X - c|) = \Sigma |x_i - c| p_i$, where $p_i = P(X = x_i)$. Prove that $g'(c) = P(X < c) - P(X > c)$ if $c \neq x_i$ for all i.

(*e*) For the situation in (*d*), show that the set where the first absolute moment is a minimum is always a closed nonempty interval (a closed interval is one which contains its endpoints, if any). Determine this interval. This interval is commonly called the **median** of the range of the random variable X. If $P(X = x_i) = 1/n$ for all i, then the interval is called the median of the set of numbers $x_1, \cdots, x_n$.

(*f*) Give an example where $c = E(X)$ does not minimize $g(c)$, that is, $E(X)$ is not in the interval.

4.8 Suppose n true dice are rolled. Let X be the total. Find $E(X)$ and $s(X)$.

4.9 Express the third moment of X about $E(X)$ in terms of the moments about zero. Express the third moment of X about zero in terms of the moments about $E(X)$ and $E(X)$ itself.

4.10 Let S be the positive integers and let $P(n) = 2^{-n}$ for each n in S. Suppose $X(n) = n$ for each n. Find $E(X)$ and $V(X)$.

4.11 *Poisson distribution.* Let $f(n) = e^{-c}(c)^n/n!$ for non-negative integers and f is zero otherwise. Show that f is a density function. Find $E(f)$ and $V(f)$.

4.12 Give an example where $E(X)$ and $E(Y)$ exist but cov (X, Y) does not. Prove that cov (X, Y) exists whenever the ranges of X and Y are bounded sets of real numbers.

4.13 The mathematics department gives a placement examination having 25 questions, with five choices for each one. Assume that for each question a student either knows the correct answer or makes a guess which has probability $\frac{1}{5}$ of being correct.

(*a*) If such a student knows the answer to n questions, what is his expected score?

(*b*) If the results of the guesses are independent, what is $s(X)$?

(*c*) What penalty for wrong answers would cause guessing to have no effect on the expected score? Compare the variance from guessing and from no guessing, with such a penalty.

(*d*) If a score of 15 is required to pass and there is the penalty for guessing as in (*c*), should the student who wants to pass guess (and at how many problems) if he knows that he knows precisely 14 correct answers? If he knows that he knows precisely 16 correct answers?

4.14 Refer to Exercise 4.13.

(*a*) If the student gets 20 answers correct, and answers all 25 questions, what is the expected number of answers that he knew?

(*b*) If the student is selected at random from a population which has probability $\frac{1}{26}$ each of knowing $0, 1, \cdots, 25$ correct answers, what is the probability that he knew 15 correct answers? What is the expected number of correct answers that he knew?

(*c*) If the student is selected at random from a population which has probability $\frac{1}{11}$ each of knowing $15, \cdots, 25$ correct answers, what is the probability that he knew 15 correct answers? What is expected number of correct answers that he knew?

(*d*) Compare the answers in parts (*b*) and (*c*) and reflect on what they imply about the question in (*a*).

5. CHEBYCHEV'S INEQUALITY AND "THE LAW OF AVERAGES"

Moments give information about the structure of a probability density. The variance in particular is a commonly used rough measure of the spread of a density function about its mean (provided the variance exists, of course). In this section we shall study the spread of the density function of one random variable and the spread of the density function of the sum of several independent random variables, as measured by their variances. We begin with one random variable.

Theorem 1. *Chebychev's Inequality.* If X is a random variable and the rth absolute moment of X is defined for some positive r, then for each real number c and each $a > 0$,

(*a*) $P(|X - c| > a) < E(|X - c|^r)/a^r$ if $P(X \neq c) > 0$.

(*b*) $P(|X - c| \geq a) \leq E(|X - c|^r)/a^r$. The second equality occurs if and only if $P(|X - c| = a) + P(X = c) = 1$.

(*c*) In particular, if $V(X)$ is defined and not zero, $P(|X - E(X)| > a) < V(X)/a^2$ and $P(|X - E(X)| \geq a) \leq V(X)/a^2$.

Proof.

$$E(|X - c|^r) = \Sigma |X(s) - c|^r P(s) = \sum_{|X(s) - c| > a} |X(s) - c|^r P(s)$$

$$+ \sum_{|X(s) - c| \leq a} |X(s) - c|^r P(s) > \sum_{|X(s) - c| > a} a^r P(s) = a^r P(|X(s) - c| > a)$$

To see that the last inequality is strict, we argue as follows. Of the two sums in the preceding expression, at least one is not zero. If the first is not zero, the replacement of $|X(s) - c|$ by a makes it strictly smaller. If the first is zero and the second is not zero, replacing it by zero makes it strictly smaller. If both are zero then $P(X \neq c) = 0$ contrary to assumption. Since neither sum is increased by the changes, and at least one of the sums is decreased, the inequality is strict. Dividing through by a^r then yields the first inequality.

The proof of the second inequality is similar. The assertion that if equality occurs, then $P(|X - c| = a) + P(X = c) = 1$ comes from dividing the first sum in the proof into three parts: $|X(s) - c|$ greater than, equal to, and less than a, respectively, and noting that the first and third sums must be zero for the equality to hold. The converse assertion is immediate.

We see from the theorem that equality in (*b*) is possible if and only if the range of X is a subset of $\{c - a, c, c + a\}$.

The tails of a probability distribution or density are the parts "far" from the mean. Although Chebychev's inequality is crude, it does give estimates of the amount of probability or "mass" in the tails.

Example 1. A coin is supposed to be true. How many times should it be tossed in order to make the probability 0.99 or more that the fraction of heads is between 0.45 and 0.55?

If the coin is tossed n times and X is the number of heads, X is binomially $(n,\frac{1}{2})$-distributed. We want $P(|X - n/2| \geq 0.05n) \leq 0.01$. (If 0.45 and 0.55 are allowed fractions of heads, the first equality should be deleted.) Using part (c) of Theorem 1, it suffices to require $V(X)/(0.05n)^2 \leq 0.01$. Since $V(X) = npq = n/4$, this yields $n \geq 10^4$. We shall later see, with the aid of the normal approximation to the binomial distribution, that $n > 430$ is actually sufficient.

Chebychev's inequality can also be used to give us information about the tails of a random variable which is the sum of random variables. We expect, for example, that if an experiment is repeated over and over, the observed frequency of an event should tend in some sense to the expected frequency. Now probability theory is a formal mathematical theory, just as plane geometry is. The theory does not refer directly to real life. But just as in plane geometry there is an intuitive correspondence intended between the objects of the theory and situations in the real world. In particular, since observed frequencies tend to limiting values as the number of trials of an experiment becomes large, we expect a theoretical counterpart to this behavior. The earliest and simplest such theoretical counterpart is given below in Theorem 2.

Theorem 2. Let $A_1, \cdots, A_n, \cdots$ be Bernoulli trials with probability p of success. Let X_i equal 0 or 1, the number of successes on the ith trial, whence $S_n = X_1 + \cdots + X_n$ is the number of successes in n trials and S_n/n is the fraction of successes in n trials. Then, given any $a > 0$, $P(|S_n - E(S_n)|/n > a) = P(|S_n/n - p| > a)$ tends to zero as n increases.

Proof. $E(S_n) = np$ and $V(S_n) = npq$. Letting $X = S_n/n$ in Theorem 1(c), we have $P(|S_n/n - p| > a) < V(S_n/n)/a^2 = pq/na^2 \leq 1/4na^2$. The last inequality follows from $pq \leq 1/4$ for all p between 0 and 1 inclusive (Exercise 5). It would have been sufficient, for purposes of this proof, to simply use the obvious estimate $pq \leq 1$. The last term would then be $1/na^2$ instead.

For each fixed a, the last expression tends to zero as n increases.

We immediately obtain the following justification of the relative frequency interpretation of probabilities.

COROLLARY 3. Let E be any event in a probability space (S,P) and then suppose we have repeated independent trials. Let F_n be the observed

fraction of the time that E occurs. Then for any real number a,

$$P(|F_n - P(E)| > a) < 1/4na^2$$

Given any real number d, if we wish $P(|F_n - P(E)| > d) > a$, it suffices to make $n \geq 1/4da^2$ trials. In particular, $P(|F_n - P(E)| > a)$ tends to zero as n tends to infinity.

Theorem 2 and Corollary 3 tells us that the relative frequency of the occurrence of an event in repeated independent trials approaches the probability of the event as the number of trials grows large. This approach is in the sense that for each individual sufficiently large n ($n \geq N$, say) the probability $P(D_n)$ of a certain deviation a on trial n of F_n from $P(E)$ is small (event D_n). However, it may conceivably happen that $P(\bigcup_{n \geq N} D_n)$ is not small. The assertion that each $P(D_n)$ is small for $n \geq N$ is a form of what is known as the **weak law of large numbers.** The stronger assertion that $P(\bigcup_{n \geq N} D_n)$ is small also can be proven for the preceding situation with the aid of nondiscrete sample spaces. (Nondiscrete sample spaces are discussed in Chapter 4.) It is a form of what is known as the **strong law of large numbers.**

The laws of large numbers are mathematical expressions of what is commonly called "the law of averages." Most nonscientists who use this phrase have a mistaken impression of what it means. Some think that after a large number of trials, the number of occurrences S_n will be quite close to the expected number of occurrences $E(S_n)$. In the terminology of Theorem 2, they believe that $P(|S_n - E(S_n)| > a)$ will be small for large n. They are surprised to learn that the reverse is generally true. For Bernoulli trials, for example, for any preassigned $a > 0$, $P(|S_n - E(S_n)| > a)$ tends to 1 as n increases. (See Feller, II. 12, Problem 20, for the case where n is even. This readily gives a proof for all n, odd or even.)

Theorem 2 does not assert that $P(D_n)$ steadily decreases as n increases. It says instead merely that $P(|S_n/n - E(S_n/n)| > a)$ ultimately becomes arbitrarily close to zero as n increases.

Others think that once a deviation of S_n from $E(S_n)$ has occurred, there is a tendency for the values S_{n+r} to return to $E(S_{n+r})$. Put simply, after twenty heads in a row, they think a tail is more probable. Thus if $X_1 = \cdots = X_{20} = 1$ in a coin toss, they expect $P(X_{21} = 1) < p$. This amounts to denying independence of trials. Experiments have shown that under the usual conditions, the assumption of independence gives a mathematical model in agreement to high accuracy with the observed results. Denial of independence (except, perhaps for very tiny and refined corrections that may be ignored) does not. Evidence of the popular belief that the outcomes of repeated coin tosses are dependent is cited at the end of Section 2.7.

There are a number of ways to establish the strong law of large numbers for Bernoulli trials. The one we present below is computationally a little tedious but is conceptually very natural and elementary.

As we saw in Examples 3.4 and 3.5, if S_n is an (n,p)-binomially distributed random variable, $E(S_n) = np$ and $E([S_n - E(S_n)]^2) = npq$. This suggests that the kth moment of S_n about the mean might be of the form $a_1 n + a_0$, where a and b are constants. If this is the case, the kth moment of S_n/n about p has the form $(a_1 n + a_0)/n^k$. This motivates the following proof of the strong law of large numbers for Bernoulli trials.

Theorem 4. Given the circumstances of Theorem 2, for each $a > 0$ there is a constant c such that $P(|S_n/n - p| > a) < c/n^2$ for all n. Consequently $P(\bigcup_{i=n}^{n+r} D_i)$ tends to zero as n increases, simultaneously for all r.

Proof. We compute the fourth moment of S_n about $E(S_n)$ by the technique of Examples 3.4 and 3.5. The idea of the computation is to express $k^4 = (k)_4 + a_3(k)_3 + a_2(k)_2 + a_1(k)_1$, where the a_i are constants to be determined. They turn out to be 6, 7, and 1, respectively. This gives

$$E(S_n^4) = (n)_4 p^4 + 6(n)_3 p^3 + 7(n)_2 p^2 + (n)_1 p$$

Now

$$E((S_n - np)^4) = E(S_n^4) - 4npE(S_n^3) + 6(np)^2E(S_n^2) - 4(np)^3E(S_n) + (np)^4$$

Thus we need to find $E(S_n^3)$ and $E(S_n^2)$. A computation similar to the preceding gives $E(S_n^3) = (n)_3p^3 + 3(n)_2p^2 + np$ and $E(S_n^2) = (n)_2p^2 + np$. Substituting and simplifying, we have

$$E((S_n - np)^4) = 3(pq)^2n^2 + (pq - 6p^2q^2)n = a_2n^2 + a_1n$$

which refutes our conjecture.* Thus $E((S_n/n - p)^4) = (a_2 n + a_1)/n^3$.

The fourth moment is the same as the fourth absolute moment. (This is not true for the third moment, which is why we did not consider using it.) Using Theorem 1(*a*) gives

$$P(|S_n/n - p| > a) < \frac{a_2 n + a_1}{a^4 n^3} \le \frac{a_1 + a_2}{a^4 n^2} = \frac{c}{n^2}$$

Now

$$P\left(\bigcup_{i=n}^{n+r} D_i\right) \le \sum_{i=n}^{n+r} P(D_i) \le c\sum_{i=n}^{\infty} i^{-2}$$

Since $\sum i^{-2} < \infty$ by the integral test, or note that

$$\sum_1^{\infty} n^{-2} = 1 + \sum_1^{\infty}(n+1)^{-2} \le 1 + \sum_1^{\infty}\frac{1}{n(n+1)} = 1 + \lim_{m\to\infty}\sum_1^{m}\left(\frac{1}{n} - \frac{1}{n+1}\right) =$$

$$1 + \lim_{m\to\infty}\left(1 - \frac{1}{m+1}\right) = 2$$

* Later, when we compute the moments for the normal approximation to the binomial distribution, we shall find the fourth moment about the mean is $3(npq)^2$ for the normal approximation. This is the leading term of our result for the fourth moment of the binomial distribution. This knowledge would have led us to a more accurate conjecture.

it follows that $\sum_{i=n}^{\infty} i^{-2}$ tends to zero as n increases, hence $P(\bigcup_{i=n}^{n+r} D_i)$ does so for all r simultaneously.

The following comparison between the weak and the strong law of large numbers may be enlightening. The weak law asserts that

$$\max_{k\ge n} P(|S_k/k - p| > a) \to 0 \text{ as } n \to \infty.$$

The strong law asserts that

$$P(\max_{k\ge n} |S_k/k - p| > a) \to 0 \text{ as } n \to \infty.$$

SUGGESTED READINGS

Feller, IX.6 (Chebychev's inequality) and IX.7 (a proof of a more general strong law of large numbers by one of Kolmogorov's inequalities).

EXERCISES

5.1 Give an example of a random variable such that the rth absolute moment is not defined for any $r > 0$. Thus there are random variables to which Theorem 1 does not apply.

5.2 $P(|X - E(X)| \ge a) = V(X)/a^2$ is possible if and only if the range of X is contained in $\{E(X) - a, E(X), E(X) + a\}$ and also $P(X = E(X) + a) = P(X = E(X) - a)$.

5.3 Give an example of a random variable with unbounded range such that the rth absolute moment is defined for every r.

5.4 If the rth absolute moment of X is defined, prove that

$$F_{|X-c|}(a) \ge 1 + f_{|X-c|}(a) - E(|X - c|^r)/a^r$$

5.5 Prove that for $0 \le p \le 1$, $p(1 - p) \le \frac{1}{4}$.

5.6 Refer to the proof of Theorem 4. Check the expansion of k^4 by the following technique.

(*a*) Prove that if two polynomials of degree n agree at $n + 1$ distinct points, they are necessarily identical.

(*b*) Verify that the expansion of k^4 is valid for the given coefficients at 5 distinct points, therefore it must be valid identically.

(*c*) Derive the coefficients of the expansion of k^4.

(*d*) Carry out the computation of $E(S_n^4)$.

(*e*) Repeat (*a*) to (*d*) for $E(S_n^3)$ and $E(S_n^2)$.

5.7 The computation of the fourth moment about the mean is long enough so there is some chance of error. It is worthwhile forming the habit of checking the results of the longer calculations without simply repeating the calculation (and very likely the errors). Verify that the moments computed in the proof of Theorem 4 satisfy the following checks.

(*a*) If $p = 0$, $S_n = 0$, therefore $E(S_n) = 0$ and the moments about zero, which is the mean, are all zero.

(*b*) If $p = 1$, $S_n = n$, therefore $E(S_n) = n$ and the kth moment of S_n about zero is n^k. The kth moment about n is zero.

(*c*) Compute $E((S_n - np)^4)$ from the formula for the cases $p = \frac{1}{2}$, $n = 0, 1, 2, 3, 4$. Also compute it directly in these cases and compare the results.

(*d*) Program a computer to carry out the calculation in (*c*) for $n = 1, \cdots, N$, and for various p. If $p = \frac{1}{2}$, up to what N can the computer go in 5 minutes? Check your computer program by comparing the results with parts (*a*) to (*c*). Also note that for $p = \frac{1}{2}$, the odd moments are zero.

**5.8 (*a*) Program a computer to calculate $E(S_n^k)$ for $k = 1, \cdots, N$.

(*b*) Extend the program to the calculation of $E((S_n - E(S_n))^k)$.

(*c*) Check the program by comparing the results with those from the previous problem.

(*d*) It is known that the even $2m$th moments about the mean of the normal approximation to the binomial distribution are $(2m)!(npq)^m/2^m m!$. This suggests the conjecture that the highest power of n in the expression for the $2m$th moment of the binomial distribution about the mean is n^m. Do your computer results support this?

(*e*) The odd moments about the mean of the normal approximation are zero, as are the odd moments for the binomial distributions with $p = \frac{1}{2}$. This suggests that the highest powers of n for odd moments of the binomial distribution "grow" less rapidly than for even moments. For even kth moments we conjectured the highest power is $k/2$. Thus for odd moments we expect it to be less than $k/2$.

Calculate the third moment about the mean by hand. Does it agree? Do the computer results for odd moments support the conjecture?

5.9 (*a*) Prove Kolmogorov's inequality: If X is a random variable with $|X(s)| \leq 1$ for all s in S, then for every $a > 0$ we have $P(|X| \geq a) \geq E(X^2) - a^2$.

(*b*) If a true coin is tossed n times, what does Kolmogorov's inequality say about the probability that the fraction of heads will satisfy $|S_n/n - \frac{1}{2}| \geq 0.05$? For what numbers n of tosses does the inequality give nontrivial information?

5.10 Consider n Bernoulli trials. Suppose a player with an initial capital of V_0 units bets a fixed fraction, f, of his current capital V_n on the outcome of each trial. With probability p on each trial he wins and receives an additional amount equal to his bet. With probability q he loses his bet. Suppose that his capital is infinitely divisible.

(*a*) If n is even and the player wins exactly $n/2$ times, prove that for all $f > 0$, $V_n < V_0$. From this, show that the probability that $V_n < V_0$ is greater than $\frac{1}{2}$ if $p = \frac{1}{2}$.

(*b*) Suppose $p = \frac{2}{3}$ and that the player wins exactly $2n/3$ times (take n to be a multiple of 3). For which f is $V_n > V_0$?

**(*c*) Prove that V_n/V_0 tends to zero with probability 1 as n increases if $p = \frac{1}{2}$ and $f > 0$.

5.11 In Example 1, we used the second moment and Theorem 1 to find that $n \geq 10^4$ sufficed. In the proof of Theorem 4, we found the fourth moment about the mean. Use this in Theorem 1(*b*) to obtain another estimate of n.

5.12 *Extension of the weak law of large numbers.* Suppose $X_1, \cdots, X_n$ are independent random variables, that $E(X_i)$ and $V(X_i)$ are defined for all i, and that there is a constant M such that $V(X_i) \leq M$ for all i. Let $S_n = X_1 + \cdots + X_n$. Then the average gain in n trials is S_n/n and the expected average gain in n trials is $E(S_n/n)$. Prove that, given any $a > 0$, $P(|S_n/n - E(S_n/n)| > a)$ tends to zero as n increases. Note that the proof of this more general theorem is easier than that given for Theorem 2 because we had to repress our urge to compute.

5.13 *Failure of the classical gambling systems.* A bet in a gambling game is a random variable. Most (but not all) of the standard gambling games consist of repeated independent trials, which means that the bets B_i are independent. Further, there is a constant K such that $|B_i| \leq K$ for all i.

(*a*) Show that there thus is a constant M such that $V(B_i) \leq M$ for all i.

(*b*) It is also true in such games that there is a minimum allowable bet. The "odds" are set so that the expected loss $E(B_i)$ is bounded above by a fixed negative number m, regardless of what bet is selected at the ith trial. Show that $E(S_n/n) \leq m$, where $S_n = B_1 + \cdots + B_n$.

(*c*) Conclude from (*a*), (*b*), and Exercise 12 that $P(S_n < 0)$ tends to one as n increases, that is, it is more and more certain that the player will be behind if he continues to play.

(*d*) Show further that if $L < 0$ is any preassigned loss, $P(S_n < L)$ tends to one as n increases, that is, it is more and more certain that the player will lose his entire capital, no matter how large, if he continues to play.

In spite of the above facts, long well known to mathematicians, thousands of man-years annually are spent in a futile search for gambling systems (ways of choosing the bets B_i) which will prevent $P(S_n < 0)$ from tending to one in games of the above type.

For an account of some of the systems that have been tried, and an account of situations where winning gambling systems are possible, see *Thorp* or *Wilson*.

5.14 Do the analog of Exercises 5.12 and 5.13 for the strong law of large numbers. *Hint:*

$$
\begin{aligned}
E([S_n - E(S_n)]^4) &= E\left(\left[\sum_{i=1}^{n} (B_i - E(B_i))\right]^4\right) \\
&= \sum_{i=1}^{n} E([B_i - E(B_i)]^4) + 6n(n-1)V(B_i) \\
&= nE([B_i - E(B_i)]^4) + 6n(n-1)V(B_i) \leq Cn^2
\end{aligned}
$$

for some constant C. Thus,

$$
E\left(\left[\frac{S_n}{n} - E\left(\frac{S_n}{n}\right)\right]^4\right) \leq \frac{C}{n^2}
$$

and Theorem 4 follows as before.

6. CARTESIAN PRODUCTS OF INFINITELY MANY DISCRETE SPACES*

The assertion of the strong law of large numbers of Theorem 5.4 makes it natural at this point to consider an infinite sequence of Bernoulli trials and to want to assert that $P(\bigcup_{i=n}^{\infty} A_i)$ tends to 0 as n increases. We previously avoided this formulation because an infinite sequence of Bernoulli trials leads us beyond discrete sample spaces. This is one of the

* This section may be omitted.

motivations for extending probability theory beyond discrete sample spaces.

If we have an infinite sequence of Bernoulli trials, the possible outcomes of the sequence are represented by the set of all possible infinite sequences of zeros (failure) and ones (success). For example, 010101 $\cdots$ represents failure on every odd trial and success on every even trial. There are uncountably many such sequences. To see this, suppose instead that there were countably many. Then we could list them all in order in a column as below, where a_{mn} represents the result of the nth trial of the mth infinite sequence a_m.

$$\begin{array}{l}
a_1 = a_{11}a_{12}a_{13} \cdots\cdots a_{1n} \cdots \\
a_2 = a_{21}a_{22}a_{23} \cdots\cdots a_{2n} \cdots \\
\vdots \qquad \vdots \qquad\qquad \vdots \\
a_m = a_{m1}a_{m2}a_{m3} \cdots\cdots a_{mn} \cdots \\
\vdots \qquad \vdots \qquad\qquad \vdots
\end{array}$$

Now consider the outcome $b = b_1b_2b_3 \cdots b_n \cdots$ defined by $b_m = 1$ if $a_{mm} = 0$ and $b_m = 0$ if $a_{mm} = 1$. Then b disagrees with each a_m in the mth place. Thus b is not in our list, contrary to the assumption that the list was complete. Thus the assumption that there are countably many outcomes is false.

This procedure can be used to show that the real numbers are uncountable (Exercise 2), that 2^S is uncountable, where S is countably infinite (Exercise 1), and that the set of all sequences of members of a set S is uncountable if S has at least two points.

The concept of infinite sequences of trials suggests the following extension of Definitions 2.7.1 and 2.7.2.

Definition 1. The cartesian product $\prod_{i\in I} S_i$ of an arbitrary collection $\{S_i\}_{i\in I}$ of sets is the set of all functions which assign to each i in I an element s_i in S_i. Such a function is designated $(s(i))_{i\in I}$ or $(s_i)_{i\in I}$. When I is the positive integers, the notation $(s_1, s_2, \cdots, s_n, \cdots)$ may be used. Let $E_i \subset S_i$, for all i in I. The cartesian product $\Pi\, E_i$ is the set of all functions $(s(i))$ such that $s(i) \in E_i$ for all i in I. These subsets of $\Pi\, S_i$ are called product sets. A product set such that $E_i = S_i$ except for at most finitely many i is a rectangle. A rectangle such that $E_i = S_i$ except for at most one E_i is denoted by $E_i{}^*$. Rectangles are often termed cylinder sets instead.

A rectangle may be written $\cap\, E_i{}^*$, where the intersection is finite and includes those $E_i{}^*$ such that $E_i \neq S_i$.

This notion of cartesian product and the other notions of this section generalize the simpler discussion for finite cartesian products of finite probability spaces already given in Section 2.7. We will restrict the following discussion to products of countably many probability spaces, although most of it carries over to products of arbitrarily many spaces.

If (S_i,P_i) is the probability space for the ith in a countable sequence of trials, $\Pi\, S_i$ represents the possible sequences of outcomes. Our problem, as in Section 2.7, is to devise a probability measure for $\Pi\, S_i$, given the probability measures P_i. The case where the trials are independent is again tractable.

It seems reasonable to impose the condition that this measure should agree with the one devised for finite products. Thus, the probability that the first n outcomes of an infinite sequence be the events $E_1, \cdots, E_n$, respectively, ought to be $\prod_{i=1}^n P_i(E_i)$, just as when only n trials in toto are planned. Thus for all rectangles $\bigcap_{i=1}^n E_i{}^*$ the product measure P is defined for independent trials by $P(\bigcap_{i=1}^n E_i{}^*) = \prod_{i=1}^n P_i(E_i)$. This much agrees with our intuition.

Our next problem is, on which other sets should the product measure be defined and how should it be defined there? It is natural to try and retain as many of the axiomatic properties as we can from the simpler case of one discrete probability space. It is evident that $P(S) = 1$ (axiom $P1$) and that $P(E) \geq 0$ (axiom $P2$) for rectangles. We would like $P4$ also to be true, that is, $P(\bigcup_{i=1}^\infty R_i) = \Sigma\, P(R_i)$ for countably infinite subpartitions of rectangles. But a countably infinite (or even a finite) union of rectangles is not in general a rectangle again. We could attempt to define P by this equality for these new sets that have arisen. If we talk of $P(\bigcup_{i=1}^\infty R_i)$, we will also want to talk of the probability of the complementary event. This event is $\bigcap_{i=1}^\infty R_i'$. Thus it is also desirable to define P on such sets, where the R_i are rectangles and the R_i' are finite unions of rectangles. For example, the strong law of large numbers is, in the context of an infinite cartesian product, the assertion that $P(\bigcup_{i\geq n} D_i{}^*)$ tends to zero as n increases, where D_n is the event $P(|S_n/n - p| > a)$. Equivalently, it is an assertion about $P(\bigcup_{i\geq n} D_i')$.

If we wish to preserve axiom $P4$ and if we wish to have P defined for certain events of great interest, we shall need to have the domain of P be a class of sets which is closed under countable unions, intersections, and complements, that is, these combinations of sets in the domain are again in the domain. Such classes of sets are very important in mathematics.

Definition 2. A sigma field of subsets of a set S is any non-empty collection Σ of subsets satisfying:

(i) $\bigcup_{i\geq 1} E_i$ is in Σ if each E_i is in Σ.
(ii) E' is in Σ if E is in Σ.
(iii) $\bigcap_{i\geq 1} E_i$ is in Σ if each E_i is in Σ.

(*a*) Condition (iii) is redundant; it follows from (i), (ii), and de Morgan's laws.

(*b*) The largest sigma field is 2^S. It contains all others. The smallest sigma field is $\{\varnothing, S\}$. It is contained in all the others.

One might wonder why we do not use 2^S as the domain of P. It turns out, for reasons which are too technical to go into here, that it is generally not possible to extend the definition of P from rectangles to all subsets of S and still have axiom *P4* satisfied. It is this that necessitates the introduction of sigma fields.

A sigma field is a subset of 2^S. Thus we can apply the Boolean operations to collections of sigma fields. In particular, the intersection of a collection of sigma fields is again a sigma field (Exercise 4). It is contained in any member of the collection so it is said to be smaller than any member of the collection.

Definition 3. If $\{E_i\}$ is a collection of sets, the intersection of all sigma fields containing $\{E_i\}$, which is the smallest such sigma field, is called the sigma field generated by $\{E_i\}$.

Since any collection of subsets is included in the sigma field 2^S, the intersection always includes 2^S and is therefore defined. It can be proven that P can be extended in one and only one way to the sigma field generated by the collection of rectangles, in such a way that axioms *P1*, *P2*, and *P4* hold. It is P with this sigma field as domain which we call **the product measure** for $\Pi\,(S_i,P_i)$. We sometimes write $P = \prod_{i=1}^{\infty} P_i$ to indicate its dependence on the P_i.

If (S,P) is the cartesian product of countably infinitely many spaces, each of which has just the two points 0 and 1, and $P_i(0) = q$, $P_i(1) = p$, $i = 1, 2, \cdots$, then (S,P) is the probability space with which we describe infinite sequences of Bernoulli trials. Each particular sequence, or point of s, has probability zero if $0 < p < 1$. To see this, let $r = \max(p,q)$. Then $r < 1$. Consider a particular sequence of outcomes $s = \{s_1, \cdots, s_n, \cdots\}$. By Exercise 5 we may write $s = \bigcap_{i\geq 1} \{s_i\}^* \subset \bigcap_{i=1}^{n} \{s_i\}^*$. Thus

$$P(s) \leq P\left(\bigcap_{i=1}^{n} \{s_i\}^*\right) = \prod_{i=1}^{n} P_i(s_i) \leq r^n$$

for all n. Since $r < 1$ and $P(s)$ is non-negative, we have $P(s) = 0$.

Example 1. A certain game has three outcomes for a player—win, loss, and tie. Suppose the game is played repeatedly until either a win or loss occurs. Let w_i, l_i, and t_i be the events win, loss, and tie, respectively on each trial i, having the probabilities w, l, and t, respectively. Let W, L, and T be the events win, loss, and tie, respectively, for the repeated game. Assume the trials are independent, and that $t < 1$. Using an infinite cartesian product model, we find $P(W)$, $P(L)$, and $P(T)$.

Let W_i be the event the first win occurs on trial i. Then the event W_i is the product set

$$t_1^* \cap \cdots \cap t_{i-1}^* \cap w_i^* \qquad \text{and} \qquad P(W_i) = t^{i-1}w$$

$$P(W) = \sum_{i\geq 1} t^{i-1}w = w/(1-t)$$

Similarly, $P(L) = l/(1-t)$. Since $W + L + T = S$, $P(W) + P(L) + P(T) = P(S)$ which yields $P(T) = 0$.

We could have shown directly that $P(T) = 0$ by repeating the argument showing that points in the product space for infinite sequences of Bernoulli trials have probability zero, provided that $0 < p < 1$.

EXERCISES

*6.1 Prove that 2^S is uncountable where S is the set of positive integers.
**6.2 Prove that the real numbers are uncountable.
6.3 The complement of the product set $\Pi\, E_i'$ is $\cup\, E_i^*$.
6.4 The intersection of a collection of sigma fields is a sigma field.
6.5 The product set $\Pi\, E_i = \cap\, E_i^*$.

7. THE POISSON DISTRIBUTION AND APPLICATIONS

In probability theory certain particular probability distributions are unusually important because they arise repeatedly, both in the theory and in the applications. One of the most important finite discrete distributions is the binomial, one of the most important infinite discrete distributions is the Poisson, and one of the most important continuous distributions is the normal (Chapter 4). They are all closely interrelated. Both the Poisson and the normal distributions may (but need not) be regarded as limiting forms of the binomial distribution.

We proceed to derive the Poisson distribution as a limiting form of the binomial distribution.

Theorem 1. For a fixed positive number c, let $np_n = c$ for $n = 1, 2, \cdots$, and let f_n be the density function for an (n,p_n)-binomially distributed random variable. Then $f(k) = \lim_{n\to\infty} f_n(k) = e^{-c}c^k/k!$ defines a function f on the non-negative integers. This function is a probability density function.

If S is the non-negative integers, and P and P_n are the probability measures determined by f and the f_n, respectively, then P_n tends to P **uniformly** as n increases. This means that, given $d > 0$, there is an integer $N(d)$ depending on d such that whenever $n > N(d)$ we have $|P_n(E) - P(E)| < d$ simultaneously for all events in E (that is, $N(d)$ does not depend on E).

Proof

$$f_n(k) \equiv b(n,k)p_n{}^k q^{n-k} = \frac{(n)_k}{k!}\frac{c^k}{n^k}\left(1 - \frac{c}{n}\right)^{n-k}$$

Now $(n)_k/n^k = (1 - 1/n) \times \cdots \times (1 - (k-1)/n)$ and since k is fixed, each factor tends to 1 as n increases. If each factor of a product tends to a limit, the limit of the product exists and is the product of the limits. Thus $(n)_k/n^k$ tends to 1 as n increases. Similarly $(1 - c/n)^{-k}$ tends to 1 as n increases. Now $(1 - c/n)^n$ tends to e^{-c} as n increases. (It is shown in calculus courses that $(1 + 1/n)^n$ tends to $e = \sum_{k \geq 0} 1/k!$ as n increases. The same proof can be used to show that $(1 - c/n)^{n/c}$ tends to e^{-1} and therefore $(1 - c/n)^n$, where $n = (n/c)c$, tends to e^{-c}.) Thus $f_n(k)$ tends to $f(k) = e^{-c}c^k/k!$ as asserted. Since $\sum_{k \geq 0} f(k) = e^{-c}\sum_{k \geq 0} c^k/k! = e^{-c}e^c = 1$, f is a probability density function.

To show that P_n tends to P uniformly as n increases, we use the following idea. For $k = 1, \cdots, K$, we can pick N so large that $|f(k) - f_n(k)|$ are each very small whenever $n > N$ and thus that $\sum_{k=0}^{K} |f(k) - f_n(k)|$ is very small whenever $n > N$. If K was properly chosen, it will turn out that the entire tails $\sum_{k > K} f(k)$ and $\sum_{k > K} f_n(k)$ of f and f_n are also very small for all $n > N$. This will allow us to conclude that $\sum_{k=0}^{\infty} |f(k) - f_n(k)| < d$ for all $n > N$, which leads to the desired result. We now formalize this by showing how, starting with a given d, we can determine $N(d)$.

Given $d > 0$, first choose $K > 0$ such that the tail of f is small: $\sum_{k > K} f(k) < d/10$. Next, choose $N_1(d), \cdots, N_K(d)$, such that $n > N_k(d)$ implies that $|f_n(k) - f(k)| < d/10K$, when $k = 1, \cdots, K$. Then if $N(d) = \max(N_1(d), \cdots, N_K(d))$, $|f_n(k) - f(k)| < d/10K$ whenever $k = 1, \cdots, K$ and $n > N(d)$.

We wish to show that for $n > N(d)$, $\sum_{k \geq 0} |f(k) - f_n(k)| < d$ for this leads to

$$|P_n(E) - P(E)| = \Big|\sum_{k \in E}(f_n(k) - f(k))\Big|$$
$$\leq \sum_{k \in E}|f_n(k) - f(k)| \leq \sum |f_n(k) - f(k)| < d$$

for all $n > N(d)$. This is the desired result. Our strategy is to write

$\sum_{k \geq 0} |f_n(k) - f(k)|$ as a finite sum of "pieces," each of which is "much" less than d, which will enable us to conclude that the entire sum is less than d, for $n > N(d)$.

We have

$$\sum_{k \geq 0} |f_n(k) - f(k)| = \sum_{k \leq K} |f_n(k) - f(k)| + \sum_{k > K} |f_n(k) - f(k)|$$

We have already seen that the first sum on the right is less than $d/10$ since each of the K terms was shown to be less than $d/10K$. To see that the second sum on the right is small, it is enough to prove that $\sum_{k > K} f_n(k)$ is small for all $n > N(d)$ because $\sum_{k > K} |f(k) - f_n(k)| \leq \sum_{k > K} f(k) + \sum_{k > K} f_n(k)$ and the first sum on the right is small because of our choice of K. Now

$$\sum_{k > K} f_n(k) + \sum_{k \leq K} f_n(k) = \sum_{k > K} f(k) + \sum_{k \leq K} f(k) = 1$$

for all n. Rearranging gives

$$\sum_{k > K} f_n(k) = \sum_{k > K} f(k) + \sum_{k \leq K} [f(k) - f_n(k)]$$
$$\leq \sum_{k > K} f(k) + \sum_{k \leq K} |f_n(k) - f(k)| < \frac{d}{10} + \frac{d}{10} = \frac{d}{5}$$

Thus

$$\sum_{k > K} |f_n(k) - f(k)| < 3d/10 \quad \text{and finally} \quad \sum_{k \geq 0} |f_n(k) - f(k)| < 2d/5 < d$$

An examination of the proof shows that $d/10$ and $d/10K$ could have been replaced by d/m and d/mK, respectively, where $m \geq 4$. This is unimportant for purposes of the proof.

When the Poisson distribution is used as an approximation to the binomial distribution, one should be sure that the approximation is sufficiently accurate. There appears to be no easy solution to the problem of what the error is. If relative error is important, Exercise 2*a* shows that this error can be arbitrarily large. If absolute error is important, then for a fixed c, it is sufficient to take n sufficiently large, as Theorem 1 shows. The errors also depend on the event E. A rough guide is to use the Poisson approximation to the binomial when $p \leq 0.1$.

An extremely wide class of phenomena are found to be described by the Poisson distribution. In some cases, it is because a priori reasoning shows that a binomial model is correct. There are also other theoretical assumptions which yield the Poisson distribution and which may be satisfied in a particular application. Sometimes it is merely established empirically that certain data are well described by the Poisson distribution.

Examples include the radioactive decay of atoms, death by horsekick in the Prussian army, the distribution of stars in the sky, testing vaccine, sampling lots of industrial items for defectives, the distribution of misprints in books, and the theory of queues (waiting lines).

Example 1. A certain city has about 365×48 emergencies per year, each of which requires the use of an ambulance for an hour. The city has 5 ambulances available at all times. (*a*) What is the probability $P(U)$ that all ambulances will be in use (event U) at a given time? (*b*) What is the probability $P(N)$ that an ambulance will be needed but none are available (event N) at a given time?

We divide the year into 365×24 hours and assume that the accidents occur independently. Then each accident has probability $p = 1/(365 \times 24)$ of occurring in a given hour. The distribution of accidents is binomially $(365 \times 48, p)$-distributed. We use the Poisson approximation to compute (*a*) the probability of 5 or more occurrences, and (*b*) the probability of 6 or more occurrences. We use $np = 2$ as the parameter.

$$P(U) = 1 - e^{-2}\sum_{k=0}^{4}\frac{2^k}{k!} = 1 - \frac{7}{e^2} \doteq 0.05$$

$$P(N) = 1 - e^{-2}\sum_{k=0}^{5}\frac{2^k}{k!} \doteq 0.01$$

Demands for the services of an ambulance are much more frequent during some hours of the week than others. The hours of interest in our case are the busiest ones. Therefore the probability of a demand for an ambulance occurring in certain hours is greater than in others. The problem can be reformulated to take this into account.

Example 2. Small items are manufactured and packaged (pins, buttons, paper cups) with the knowledge that the probability of each being defective is 0.001. The manufacturer guarantees at least 100 good items per package. How many items should he include in each package to have the probability at least 0.99 that he meets his guarantee?

Let $100 + x$ be the number of items he includes. The distribution of defectives is $(100 + x, 0.001)$ binomially distributed. Using the Poisson approximation, we want $f(x) = e^{-c}(1 + c + c^2/2! + \cdots + c^x/x!)$ to be greater than or equal to 0.99, where $c = (100 + x)p = 0.1 + 0.001x$. The log scales of a slide rule lead quickly to $f(0) \doteq e^{-1/10} \doteq 0.905$; $f(1) \doteq 1.1e^{-1/10} \doteq 0.995$. Therefore, 101 items in each package will suffice.

Example 3. *Stars.* A certain small region of the sky is examined telescopically and found to contain a great many stars. If the region is small compared to gross irregularities, the stars will appear to be

"randomly" distributed. The Poisson approximation may be used to test the hypothesis that the stars are in fact randomly distributed.

Suppose a photograph of a small region is taken and all stars which appear to exceed a certain threshhold of intensity are counted and that there are n stars. Suppose the photograph is divided by a grid or screen of horizontal and vertical lines into m congruent squares. Consider a specified square. The assumption that the stars are "randomly" distributed should mean that the probability that a given star will be in this (or any other) square is $1/m$. Randomness should also mean that whether or not one star appears in this square is independent of whether or not any other star appears in this square. Thus the number of stars in the square is binomially $(n,1/m)$-distributed. For a fixed n/m and a sufficiently large n, the Poisson approximation should apply with great accuracy.

When the number of squares with $0, 1, 2, \cdots$ stars is tallied, it is found that the Poisson distribution with $c = n/m$ **does not** fit the data. There are too many squares with few stars and too many squares with many stars. The stars exhibit a clumping together. The explanation is that a very large fraction of stars in the sky are double stars or of still higher multiplicity.

This fact about stars was not discovered with the aid of the Poisson distribution but it is an interesting fact that it could have been discovered sooner with the Poisson distribution, if anyone had thought of using it for the purpose. (Photos are not necessary, only helpful. An accurately drawn star map would serve as well.)

There are many situations which are similar to the star example, and where the Poisson distribution does in general give a very accurate description of the data. The distribution of bacteria on a Petri plate, or of raisins in cookies, are examples. The raisins correspond to the stars and the division into cookies of the dough corresponds to the division of the photo into squares. The Poisson distribution is useful in studying statistical sampling processes of many kinds.

Suppose, for instance, that a person has m cubic centimeters of blood and that his blood contains a total of n particles of a certain type (red or white corpuscles, a certain strain of bacteria, antibodies, etc.). If a one cubic centimeter sample is extracted, the number of particles in the sample is described by a Poisson distribution with parameter n/m.

Example 4. *The radioactive decay of atoms.* In a certain fixed time interval, each atom of a radioactive substance is hypothesized to have a certain probability of decay. It is further supposed that the decay probability is the same for each atom, decay of different atoms is independent, the decay probability is unaffected by the passage of time, and that it is

Table 1. Radioactive decay data

k	$n(k)$	$nf(k)$
0	57	54.399
1	203	210.523
2	383	407.361
3	525	525.496
4	532	508.418
5	408	393.515
6	273	253.817
7	139	140.325
8	45	67.882
9	27	29.189
10	16	17.075
	2608	2608.000

quite insensitive to changes in environment (such as temperature and pressure). If there are n atoms to begin with and a probability p that a given atom will decay (more precisely, p should be the probability that the instruments will detect the decay) in the chosen time interval, then the number of atoms which decay should be an (n,p) binomially distributed random variable.

In a famous experiment by Rutherford and Geiger, the number of alpha particles detected in 2608 periods of 7.5 seconds each was recorded. Let k be the number of particles observed in a period, and $n(k)$ be the number of periods with this many observed particles. Since $\Sigma\, kn(k)/\Sigma\, n(k) = 10086/2608 \doteq 3.87$, we choose $np = 3.87$ as the parameter of a Poisson distribution ($n = 10086$, the number of detected decays, and $p = 1/2608$, the reciprocal of the number of time periods) to fit the data. If f is this Poisson distribution, Table 1 shows the results. The agreement is striking.

Example 5. Bortkiewicz determined the number of soldiers in ten cavalry troops of the Prussian army who died from a horse kick over a twenty-year period. The number of men killed in one corps in one year was tabulated. There were 200 observations. The results were as in Table 2. Again we have taken the Poisson parameter as $\Sigma\, kn(k)/n \doteq 0.61$.

Table 2. Death from horsekick in the Prussian army

Number of deaths k	0	1	2	3	4
Number of observations $n(k)$	109	65	22	3	1
Poisson fit $nf(k)$	108.8	66.2	20.2	4.2	0.6

SUGGESTED READINGS

Parzen, Sections 3.3 and 6.3. Further discussion of the Poisson distribution and examples of applications.

Feller, Sections VI.5, 6, 7, and selected exercises from VI.10. Further discussion of the Poisson distribution and applications.

EXERCISES

7.1 Show that the rth absolute moment of the Poisson distribution is defined for every non-negative r.

7.2 (*a*) Despite the fact that $|P_n(E) - P(E)| \to 0$ uniformly in E, as n increases, show that the relative error in approximating the binomial distribution by the Poisson distribution, $(P(E) - P_n(E))/P_n(E)$, can be arbitrarily large. (*Hint:* Consider elementary events.)

(*b*) Prove or disprove the assertion that the relative error of greatest magnitude for a given n and p, will always be taken on for an elementary event.

7.3 Show that the mean and variance of a Poisson distributed random variable with parameter c are both c.

7.4 The Poisson distribution was obtained from the binomial by letting $np_n = c$ and $n \to \infty$. It is tempting to suppose that the moments of the binomial distribution tend to the corresponding moments of the Poisson under these circumstances.

(*a*) Verify that this is true for the first two moments, by comparing the results of the preceding exercise with the result of taking limits of the first two moments of S_n as given in the proof of Theorem 5.4.

(*b*) What do you guess are the third and fourth moments of the Poisson distribution about zero? About the mean?

**(*c*) Prove (or disprove) that the limit of a binomial moment yields the corresponding Poisson moment.

7.5 A gambler always bets the number 7 at Nevada roulette (38 equally likely numbers). He finds that in 38 trials his number comes up three times. What is the à priori probability that he would have been at least this fortunate?

7.6 *A computer study of the error in the Poisson approximation.* Program a computer to determine for several values of c the n values (interpolating between integers when necessary) for which the maximum difference between any term of the Poisson distribution and the corresponding term of the binomial distribution is less than $d = 0.01$. Do the same for other selected values of d and then draw curves of constant d in the $n - c$ plane, where n is a binomial parameter and c is the Poisson parameter. This gives a pictorial description of the limits for the maximum possible error.

7.7 In a group of 1000 people, what is the probability that precisely one was born on Christmas? Precisely k? State any assumptions you make.

7.8 A do-it-yourself boat kit needs 4000 screws of a certain size. The probability that a screw is defective is 0.0001. How many screws should the manufacturer pack so that the probability is at least 0.9 that there is enough? So that the probability is 0.99? 0.999?

7.9 A book of n pages contains an average of m misprints per page.

(*a*) Estimate the probability that a given page will contain more than k misprints.

(*b*) Estimate the probability that at least one page will contain more than k misprints.

State any simplifying assumptions that you make.

7.10 Suppose Example 1 is reformulated to take into account the fact that the probability of a demand for an ambulance is greater for some hours than for others, with average overall demand the same. How would you expect the values of $P(U)$ and $P(L)$ computed on this basis to compare with those found in Example 1? Give intuitive arguments to support your answer.

7.11 We define the exponential function e^x by $e^x = \sum_{n=0}^{\infty} x^n/n!$. In particular the base of natural logarithms, e, is defined as $\sum_{n=0}^{\infty} 1/n!$.

(*a*) Prove that the series for e^x converges for all real x.

(*b*) Prove that $\lim_{n\to\infty} (1 + 1/n)^n = e$. *Hint*: Proceed as follows.

(i) Expand $(1 + 1/n)^n$ by the binomial theorem and observe that $(1 + 1/n)^n < e_n = \sum_{k=0}^{n} 1/k! < e$ for each n.

(ii) Observe that, given $\epsilon > 0$ and a partial sum e_m, then for each fixed k and all sufficiently large n,

$$(1 + 1/n)^n = \sum_{k=0}^{n} [(1 - 1/n)\cdots(1 - (k - 1)/n)]/k! \geq e_m - \epsilon$$

(iii) From (i) and (ii) it follows that for all sufficiently large n,

$$e_m \leq (1 + 1/n)^n < e.$$

Taking the limit as n tends to infinity then yields

$$\lim_{n\to\infty} (1 + 1/n)^n = e$$

CHAPTER 4

Continuous Probability

1. INTRODUCTION

Think of the set of real numbers as represented by the points on a line—the real line. The rational numbers are a subset of this line, but a subset that is riddled with tiny holes everywhere. In these holes belong numbers which are not rational—not the quotient of integers, and therefore called irrational. As we have seen in Exercise 3.6.2, there are many more irrationals than rationals in the sense that the rationals are countably infinite but the reals, and hence the irrationals, are not.

Rational numbers are precisely those numbers which are repeating decimals (Exercise 1). It is quite easy to manufacture decimals which are not repeating. Consider $0.010010001\cdots$, for example. Yet any such number t has rational numbers which are arbitrarily close to it. We simply choose the rational r_n which has the same decimal expansion in the first n decimal places and zeros afterwards. Then $|t - r_n| < 10^{-n}$. Thus the sequence $\{r_n\}$ of rationals approaches t as n increases.

When we add all other decimals to the rationals, it turns out that there are no "holes" left. Every sequence of numbers with terms which all get "close" as n increases have some decimal number as a limiting value. In this sense the real numbers are "complete." As we proceed along the real line we go "continuously," without jumping over holes. The real numbers are called the continuum.

The notion of continuity in mathematics is much the same. A one-dimensional line with no gaps or breaks, but possibly very kinky, is "continuous" (it might be thought of as a copy of the real line which has been bent and twisted). A function with a graph in the plane like this is called continuous. A sheet or plane with no rips or cracks is a "continuous" two-dimensional object. The cartesian product of n copies of the real line (with a suitable notion of closeness or, equivalently, of distance) is an n dimensional "continuous" object. Suitable subsets of the above examples, such as intervals on the line, interiors of triangles in the plane, and solids and hypersolids in three and more dimensions, also qualify.

Definition 1. A finite interval I is a subset of the real line having the form $I = \{t : a \leq t \leq b\}$, or this form with one or both equalities removed. An infinite interval has one or both of the a or b restrictions removed. An interval is open if every point of the interval contains points of the interval on both sides of it. It is closed if it contains its endpoints (if any). A finite interval is normally written (a,b) with a round bracket to show the endpoint is missing and a square bracket to show that it is included. For example, $[a,b) = \{t : a \leq t < b\}$. If the a restrcition is removed, the left bracket is round and a is replaced by $-\infty$. If the b restriction is removed, the right bracket is round and b is replaced by ∞. An interval which contains more than one point is called nontrivial.

(*a*) The real line R is the only interval which is both open and closed. $R = (-\infty, \infty)$.

(*b*) If $a > b$, the interval is empty. An interval contains precisely one point if and only if it is closed and $a = b$. If it is not closed and $a = b$, it is empty.

(*c*) If an interval contains more than one point, there is a 1 to 1 correspondence between it and the entire real line (Exercise 3). Thus a nontrivial interval is not countable. This occurs if and only if $a < b$.

The real numbers are perhaps the most widely used construction in mathematics. In particular, the range of many (perhaps most) random variables of interest is the set of real numbers or an interval thereof. This leads us to the consideration of uncountable (and therefore nondiscrete) sample spaces.

Example 1. A series of projectiles is fired from a gun. The speed s of each projectile as measured by a certain instrument at a certain point is recorded. This speed s is the outcome of an experiment and it seems reasonable to suppose that it can have any value in some interval of real numbers.

One might argue that the measuring instrument only supplies values of s to a certain number of decimal places, therefore there is no need to go beyond finite sample spaces. Of course, we might want to discuss the possible values which can be recorded by all conceivable measuring instruments. This would lead us to allowing all terminating decimals (ones which are ultimately 0 in each place) as outcomes. These are a subset of the rational numbers, therefore countable, so the space is discrete.

There are two reasons for introducing concepts of continuity in applied mathematics. First, the best model might be continuous, rather than

discrete. Second, even if the correct model is discrete, a continuous one may lead to great simplicity of formulation, or may be much easier to handle mathematically. The errors due to the approximation of a discrete situation by a continuous one will generally be so much smaller than the smallest experimental errors as to be inconsequential.

Example 2. Buffon's needle experiment. A smooth plane surface is marked with parallel lines spaced by a distance d. A needle of length $L < D$ is dropped "at random" on the surface. What is the probability that the needle will touch one of the lines?

Suppose the ends of the needle can be distinguished (one might be colored), and that a coordinate system is used. Let the x-axis be one of the parallel lines. Then we can describe an outcome by the coordinates (x,y) of a specified endpoint and the angle θ between $0°$ and $360°$, measured counterclockwise from the positive x-axis to the direction of the needle. Whether the needle touches a line is not affected by the value of x. Also, "random" should mean that the situation is the same for each of the horizontal strips bounded by two adjacent parallel lines. (Or, we could restrict our attention to one strip and do the experiment over again whenever (x,y) is not in that strip.) Thus we may as well describe outcomes by the pair of numbers (y,θ), where $0 \leq y \leq D$ and $0 \leq \theta < 2\pi$ (radians), and the sample space is the rectangle in Figure 2.

In order that the needle touch a line, we must have either $y + L \sin \theta \geq D$ or $y + L \sin \theta \leq 0$. These regions are shaded in Figure 2. Now randomness seems as though it should mean that all values of y, $0 \leq y \leq D$, are "equally likely" in some sense. In particular, we would expect that the probability of a value of y falling in a certain interval $[a,b]$ would be

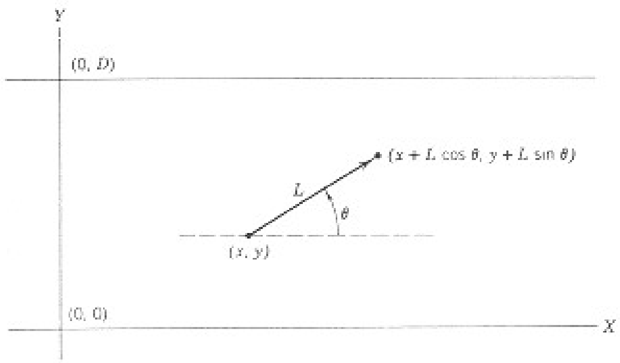


Figure 1. Buffon's needle experiment.

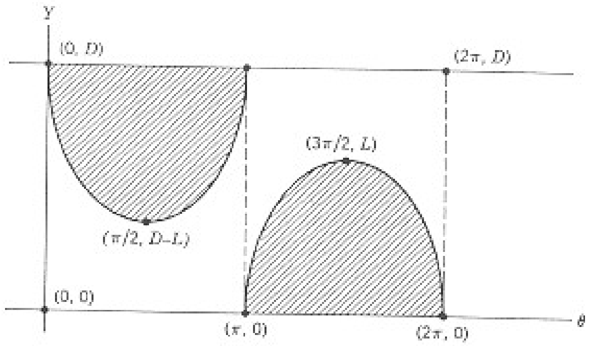


Figure 2. The shaded areas show the (y,θ) pairs for which the needle touches one of the two horizontal lines in Figure 1.

proportional to the length of this interval, that is, $P(a \le y \le b) = (b - a)/D$. Similarly, we would expect that no direction be preferred, and in particular that $P(c \le \theta \le d) = (d - c)/2\pi$. The y coordinate and the angle should be independent so we expect $P(a \le y \le b, c \le \theta \le d) = (b - a)(d - c)/2\pi D$. Thus the probability of a rectangle is equal to the fraction of the total area which the rectangle occupies. The shaded region can be approximated as closely as we like by rectangles. It is plausible now to argue that the probability that an outcome will be in the shaded region is the fraction of the total area which the shaded region occupies.

The area of the shaded region is $\int_0^{\pi} (L \sin \theta)\, d\theta - \int_{\pi}^{2\pi} L \sin \theta\, d\theta$ or $4L$. The area of the rectangle is $2\pi D$. Thus the probability would be, by this argument, $2L/\pi D$. The results of experiments with needles verify this classical formula. We appear to have a probabilistic way of computing π.

Many of the concepts that came from our "natural" solution to the probability problem in Buffon's needle experiment can be used as a basis for extending our probability notions to nondiscrete sample spaces. We shall consider this in the next section. Before we do so, however, it will be instructive to consider one more example.

Example 3. Bertrand's "paradox." A chord is chosen "at random" in a circle of radius R. What is the probability that it is shorter than $R\sqrt{3}$, the length of the side of an inscribed equilateral triangle?

(*a*) Suppose we fix one end P of the chord on the circle and choose the other endpoint Q "at random." If this means choosing the angle between the chord and the tangent to the circle at P "at random," then we see from Figure 3 that the chord is shorter if $0 \le \theta < \pi/3$ or if $2\pi/3 < \theta \le \pi$.

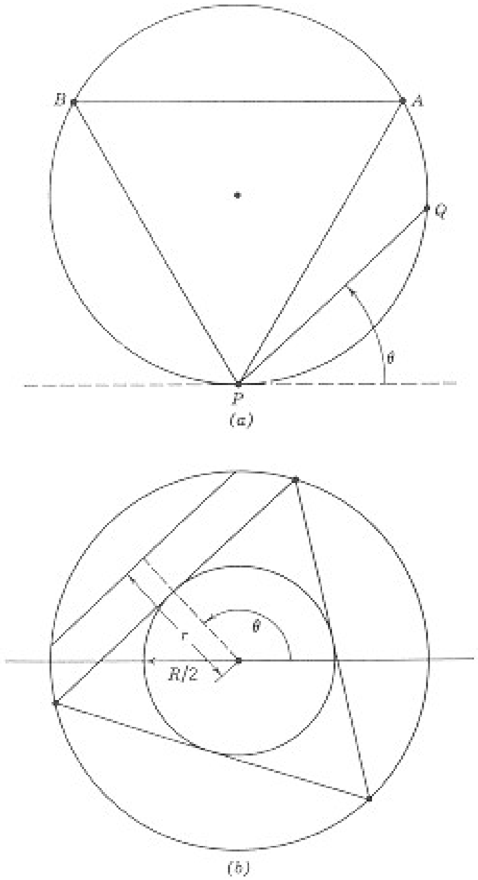


Figure 3. Bertrand's "paradox."

(If chords of zero length are forbidden, delete the equalities. The solution is unchanged.) Interpreting "at random" to mean, as before, that the probability of a θ interval is proportional to its length, we have the desired probability as $\frac{2}{3}$.

If instead we choose the position of Q on the circle "at random," the chord is shorter if Q is between P and A or between P and B. This is $\frac{2}{3}$ of the perimeter of the circle so the probability is again $\frac{2}{3}$.

Note that all chords are counted twice in the above procedures. This does not affect the results.

(*b*) Suppose instead that we identify a chord by the polar coordinates (r,θ) of its midpoint. Then there is a 1 to 1 correspondence between the number pairs (r,θ), with $0 \leq r \leq R$ and $0 \leq \theta < 2\pi$, and the chords. ($R = r$ means the chord has zero length. Excluding these chords by imposing the restriction $r < R$ does not affect the result.)

If we choose the chord at random, we may argue that all r values and all θ values in the allowed intervals are, respectively, equally likely. Further, the values of r and the values of θ should be "independent." The chord is shorter than $R\sqrt{3}$ if and only if $r > R/2$ (Figure 3*b*). Taking a ratio of areas gives the probability as $\frac{1}{2}$.

(*c*) Choose the midpoint of the chord "at random." The chord is shorter than $R\sqrt{3}$ if the midpoint is outside the concentric circle of radius $R/2$. The probability of a midpoint being chosen from any region is proportional to the area of the region. Thus the probability is $\frac{3}{4}$ (area of outer ring divided by πR^2) that the chord is shorter than $R\sqrt{3}$.

The explanation for these different results is simply that there are many ways to choose a sample space and to define probabilities for the "experiment" of choosing a chord at random. The preceding illustrate three of these ways. Each of these preceding assumptions show us how to assign probabilities and are the proper assumptions for different experiments.

In (*a*), for example, P could be fixed arbitrarily. Then the circle perimeter could be divided into n equally long pieces and then one could be selected for the choice of Q. If the choice is unambiguous (that is, for all locations in the interval, the chord is either longer or shorter), the experiment terminates. If the choice is ambiguous, the interval is subdivided into n equal parts and one is chosen in the same way from this new subdivision. This continues until a decision is reached. Many repetitions of this process should lead to a relative frequency approaching $\frac{2}{3}$.

Similar experiments can be formulated for (*b*) and (*c*).

EXERCISES

1.1 Prove that a number is rational if and only if its decimal expansion (or its expansion in any base) is repeating. A number with a repeating decimal expansion is one of the form $a_1 \cdots a_k \cdot b_1 \cdots b_m c_1 \cdots c_n c_1 \cdots c_n \cdots$. The number is said to have period n. Thus a number with a repeating expansion in one base has a repeating expansion in every base. Is the period the same for every base?

1.2 (*a*) Prove that $\sqrt{2}$ is not rational. (*Hint:* Suppose that $\sqrt{2} = m/n$ is in lowest terms.)

(*b*) Prove that $m^{1/n}$ is either an integer or is irrational, where m and n are positive integers. (*Hint:* Express m uniquely as a product of powers of increasing primes via the prime factorization theorem.)

1.3 (*a*) There is a 1 to 1 correspondence between the interval $(-\pi,\pi)$ and the entire real line, therefore they have the same "number" of points. (*Hint:* Consider $f(t) = \tan^{-1} t$, where $\tan^{-1}$ is given its principal determination.)

(*b*) There is a 1 to 1 correspondence between the real line and any nontrivial interval.

1.4 A point is chosen at random in a circle of radius R. A chord is then drawn through this point in a random direction. What is the probability that it is shorter than $R\sqrt{3}$? (*Hint:* The integral which arises can be computed by using integration by parts.)

Notice that if we think of the outcomes as the result of a two step process: first, choose point (r,θ), $0 \leq r \leq R$, $0 \leq \theta < 2\pi$, second, choose angle ϕ, $0 \leq \phi < \pi$, then the 3-tuples (r,θ,ϕ) are in 1 to 1 correspondence with the outcomes. If, however, we think of the outcomes simply as the final chords which result, many 3-tuples correspond to the same outcome. Compare the situation when two dice are rolled and we are interested in the total.

1.5 In Buffon's needle experiment, determine the probability that the needle will touch a line if $D < L$.

2. BOREL SETS ON THE LINE AND CONTINUOUS DENSITY FUNCTIONS

Buffon's needle experiment and Bertrand's example pose the problem of assigning probabilities to events which are subsets of a continuum. Let us examine the problem in its simplest form.

Suppose a point t is chosen "at random" from the **unit interval** $I = \{t: 0 \leq t \leq 1\}$. What is the probability $P(E)$ that for some specified subset $E \subset I$, t will be in E? The problem is like that of the one in Section 3.6 where we assigned probabilities to an infinite cartesian product of discrete spaces. There we had certain sets—the rectangles—to which an assignment of probabilities was made that agreed with our previous notions for finitely many trials. Here it seems natural to assign to an interval $E \subset I$ a probability equal to the length $b - a$ of E, divided by the length of I. Thus $P(E) = b - a$ for intervals contained in the unit interval.

Again, we have the problem of having P satisfy axioms $P1$, $P2$, and $P4$. And again, the answer is precisely the same. There is a sigma field of subsets generated by the intervals. There is a unique P defined on this sigma field which satisfies $P1$, $P2$, and $P4$, and which assigns to each interval its length. Furthermore, there is no P which is defined on 2^I and has all these properties. Thus sigma fields need to be introduced.

*The reason that the answer is the same for I with this P as it was for

* The next three paragraphs depend on Section 3.6 and may be omitted.

countably infinite cartesian products is no accident. If we take the countably infinite cartesian product (S,Q) for Bernoulli trials with $p = 1/2$, the probability space we get is the same as (I,P) in the following sense. There is a correspondence f between S and I, which is 1 to 1 when we remove a certain countable subset of S having measure zero, such that E is in the sigma field which is the domain of P if and only if $f(E)$ is in the sigma field which is the domain of Q. Also, $P(E) = Q(f(E))$ for all such E. Thus, from the probability standpoint, the two spaces are indistinguishable. Any probability statement about one of them translates via f or f^{-1} into an equivalent statement about the other.

The correspondence f is the following. Each element $s = \{s_1, s_2, \cdots\}$ of S is a sequence of zeros and ones. Let $f(s)$ be the binary element (base two "decimal") $.s_1s_2 \cdots = \sum s_i/2^i$. Those numbers in I (except for 1) which have a terminating binary expansion have two elements of S mapped onto them. For instance, $01111 \cdots$ and $10000 \cdots$ both map onto $0.10000 \cdots = \frac{1}{2}$. The set of sequences N which are ultimately 1 in each place are a countable set (Exercise 1) and therefore are a null set (Exercise 2). The correspondence f is biunique between S-N and I.

It can be shown that intervals in I of the form $[m/2^k, (m+1)/2^k]$ correspond to certain rectangles in S-N under f^{-1} and that Q of the interval is P of the rectangle. From this and the uniqueness of P and Q on their domains it can be shown that $f \circ P = Q$ and $f^{-1} \circ Q = P$.

Definition 1. The members of the sigma field generated by the subintervals of any interval of the real line are called the Borel subsets of that interval. If I is any interval, the unique function defined on the Borel sets, satisfying $P1$, $P2$, and $P4$, and assigning to each interval a probability proportional to its length, is the uniform, or Borel, probability measure on I.

The concepts of Definition 1 extend to sets in two and more dimensions. Though we won't make use of them, we remark that they give the assignments of probabilities arrived at in the last section: probabilities are proportional to length, area or volume.

The appearance of a sigma field which is a proper subset of 2^S leads to the following generalization of our previous definition of probability measure and probability space.

Definition 2. A probability measure is a real-valued function defined on a sigma field Σ of subsets of a subset S and satisfying the axioms $P1$, $P2$, and $P4$. A probability space (S,Σ,P) is a set S, a sigma field Σ, and a probability measure with domain Σ.

Situations are easy to imagine in which the probabilities of various subintervals should not be equally weighted. Suppose, for instance, that a point is chosen at random in a circle of radius R and that we are interested in the probability that the point will lie at a distance r from the center of the circle. The probability that the point lies at a distance between r and $r + h$, where h is very small, is approximately $2\pi rh/\pi R^2$ or $2rh/R^2$. Thus the probability of a small interval of length h in the real number interval $[0,R]$ depends on its location in the interval. The interval has a "weight" proportional to r. We have $P(a < r < b) = \int_a^b (2r/R^2)\,dr$.

The function $2r/R^2 = f(r)$ suggests the notion of a continuous density function, to play a role for intervals of real numbers analogous to that played for countable subsets by discrete density functions. The other analogues of our earlier definitions are natural consequences of the introduction of sigma fields.

Definition 3. A non-negative function f which is defined on the real line, is continuous in every bounded interval except for a finite number of points, and such that

$\int_{-\infty}^{\infty} f(t)\,dt = 1$, is a piecewise continuous density function. The probability measure P_f on the real line which is associated with f is defined by $P_f(E) = \int_a^b f(t)\,dt$ for intervals E with endpoints a and b. If (S,Σ,P) is a probability space, a random variable X is any real-valued function on S such that for each Borel set F, $X^{-1}(F) \in \Sigma$. Thus for random variables $P(X^{-1}(F))$ is defined for all Borel sets F. In particular the distribution function F_X for X is defined by $F_X(t) = P(X^{-1}((-\infty,t]))$. If there is a piecewise continuous density function f such that $F_X(t) = \int_{-\infty}^{t} f(s)\,ds$, X is called a continuous random variable and f is the density function for X. The distribution function F_X for X is then a continuous function and we say that X has a continuous distribution function. The density function for X is denoted by f_X.

(*a*) $P(S) = \int_{-\infty}^{\infty} f(t)\,dt = 1$ so $P1$ holds.

(*b*) The integral of a non-negative function over an interval is non-negative so $P2$ holds.

It is shown in advanced courses that $P(E)$ extends in a unique way to all Borel sets, just as the Borel measure on the unit interval does. Also, a class of functions which is much more general than the ones that are piecewise continuous can be used. We restrict ourselves to these because they are integrable by the methods available in the first calculus course, and because many applications of interest require only this type of density function.

The uniform measure on a nontrivial interval E with endpoints a and b is the special case where $f = C_E/(b - a)$, the characteristic function of E divided by the length of E. There are discontinuities of f at a and b.

Definition 4. The **mean** or **expectation** of a continuous random variable is $E(X) = \int_{-\infty}^{\infty} tf_X(t)\,dt$, provided the integral is absolutely convergent. If g is a piecewise continuous real-valued function defined for all real numbers, the mean or expectation of the random variable $g(X)$ is defined as $E(g(X)) = \int_{-\infty}^{\infty} g(t)f_X(t)\,dt$, provided the integral is absolutely convergent. The rth absolute moment of a continuous random variable about the point c is denoted $E(|X - c|^r)$ and defined to be

$$\int_{-\infty}^{\infty} |t - c|^r f_X(t)\,dt.$$

The kth moment about c is $E((X - c)^k) = \int_{-\infty}^{\infty} (t - c)^k f_X(t)\,dt$. $V(X)$ and $s(X)$ are defined as before.

The continuous analogues of the facts established in the discrete case are true without exception. The reader should verify these (except the proofs of facts involving two or more random variables, such as $E(X + Y) = E(X) + E(Y)$, require more sophistication and so we omit them here). The analogues the reader should state and prove are listed in Exercise 8.

Numerical-valued random phenomena are described by random variables. In studying a random variable X, we are generally interested in $P(X(s) \in E) = P(X^{-1}(E))$, where E is a Borel set of real numbers, rather than in the underlying probability space (S,Σ,P). If f_X and F_X are the density and distribution functions for X, we can determine $P(X^{-1}(E))$ from them. Thus it is common to talk of the density or distribution function of a random variable, rather than the random variable itself. Reference to an underlying probability space is often abandoned. The density or distribution function is called the **probability law of X**. Certain constants used in the definition of the density function may be chosen to fit the application, like p was for Bernoulli trials. These are called the **parameters** of the probability law.

Certain probability laws are of particular importance. We give some examples and use them to illustrate the preceding definitions.

Example 1. The **uniform** probability law over (a,b) is defined by $f(t) = 1/(b - a)$, $a < t < b$, and $f(t) = 0$ otherwise. The kth moment about c of f is

$$\int_{-\infty}^{\infty} (t - c)^k f(t)\,dt = (b - a)^{-1}\int_a^b (t - c)^k\,dt$$
$$= \frac{\{(b - c)^{k+1} - (a - c)^{k+1}\}}{(k + 1)(b - a)}$$

In particular, $E(f) = (a + b)/2$ and $V(f) = E((f - E(f))^2) = (b - a)^2/12$.

The discussion is unchanged if the interval contains either or both endpoints because a piecewise continuous density function always assigns probability zero to a countable set. Thus the probability measure for Borel sets obtained from a piecewise continuous density function is unaffected if we alter the function on a countable set, or if we alter the set whose probability we are computing by adding or removing countably many points (Exercise 3).

The formula for $E(f)$ is well known to physicists as the center of gravity of a rod of unit mass with endpoints a and b. The result for $V(f)$ is the moment of inertia of such a rod about its center of gravity.

A piecewise continuous density function can be thought of as describing a unit mass which is distributed without lumps in the form of a (possibly infinite) wire of varying physical density. Discontinuities in the density function might be thought of as denoting a junction of two different kinds of wire. Then $E(f)$ is the center of gravity of such a wire and $V(f)$ is the moment of inertia of the wire about its center of gravity.

Example 2. The **Cauchy** distribution with parameters a and b. Let $f(t) = 1/\pi b(1 + (t - a)^2/b^2)$, for all t. This distribution is realized by the following physical model. Let radioactive material be placed at (a,c), $c > 0$, in the xy-plane. Suppose the x-axis is a long thin wire-like detector of decay products. Then if we assume that all angles of emission are equally likely, and that the height of the detector can be neglected, the probability that a particle arrives at an interval (more generally, any Borel set) on the x-axis is described by the Cauchy distribution (Exercise 4).

In experiments where rows of seeds are planted in a field and exposed to radiation from a "point" source of radioactive material, these conditions are realized quite accurately.

To verify that f is a probability density, we have

$$\int_{-\infty}^{\infty} f(t)\,dt = \int_{-\infty}^{\infty} \frac{dt}{\pi b(1 + (t - a)^2/b^2)}$$
$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{du}{1 + u^2} = \frac{1}{\pi}\tan^{-1} u\Big|_{-\infty}^{\infty} = 1$$

where we have made the change of variable $u = (t - a)/b$.

The expectation of f about a is defined only if $I = \int_{-\infty}^{\infty} |t - a| f(t)\, dt$ is convergent. We have

$$\int_{-\infty}^{\infty} |t - a| f(t)\, dt = -\int_{-\infty}^{a} (t - a) f(t)\, dt + \int_{a}^{\infty} (t - a) f(t)\, dt$$

The change of variable determined from the equation $v - a = -(t - a)$ yields $-\int_{-\infty}^{a} (t - a) f(t)\, dt = -\int_{\infty}^{a} (v - a) f(v)\, dv$ and replacing v by t (arbitrary changes of dummy variables are permitted with integrals, just as with summations), and reversing the limits and changing sign, yields the second integral. Thus

$$I = 2\int_{a}^{\infty} \frac{(t - a)\, dt}{\pi b(1 + (t - a)^2/b^2)} = \frac{b}{\pi} \int_{0}^{\infty} \frac{2u\, du}{1 + u^2} = \frac{b}{\pi} \log(1 + u^2)\Big|_{0}^{\infty} = \infty$$

Thus the integral is not absolutely convergent and the expectation is not defined.

The details of the discussion of absolute convergence of integrals is one of the few places where the details for piecewise continuous density functions are different than for discrete density functions. Since this concept arises in the definition of moments, we discuss it briefly. The integral $I = \int f(t)\, dt$ of a real-valued function f of a real variable is **absolutely convergent** over the range of integration, if and only if the positive part $I^+ = \int f^+(t)\, dt$ of the integral and the negative part $I^- = \int f^-(t)\, dt$ of the integral are both finite, where $f^+(t) = \max(f(t), 0)$ is the **positive part** of the function and $f^-(t) = \max(-f(t), 0)$ is the negative part of the function. It is easy to see that $f = f^+ - f^-$ and $|f| = f^+ + f^-$ (compare Exercise 3.3.9). This is completely analogous to the situation with series, where a series converges absolutely if and only if the series of positive terms converges and the series of negative terms converges. In fact, for any series $\Sigma\, a(n)$, where the terms $a(n)$ are the values of a real-valued function defined on the integers, we can define $a^+(n) = \max(a(n), 0)$ and $a^-(n) = \max(-a(n), 0)$ and proceed as above.

If $I^+ = \infty$ and $I^- < \infty$, we say $I = \infty$ or I diverges to ∞. If $I^+ < \infty$ and $I^- = \infty$, we say $I = -\infty$ or I diverges to $-\infty$. If $I^+ = \infty$ and $I^- = \infty$, the discussion is more complicated. There are a variety of procedures for attempting to "integrate" such functions which still give the usual value $\int f^+ - \int f^-$ for absolutely convergent integrals but also give finite values for certain other integrals.

A well-known example of such a procedure is the Cauchy principal value. Consider the function $f(t) = 1/t$ for $|t| > 1$ and $f(t) = 0$ otherwise.

Then

$$\int_{-\infty}^{\infty} |f(t)|\, dt = 2\int_{1}^{\infty} dt/t = 2 \log t\Big|_{1}^{\infty} = \infty$$

Thus the integral is not absolutely convergent. But $\int_{-A}^{A} dt/t = 0$ for every real A. Thus $\lim_{A\to\infty} \int_{-A}^{A} dt/t = 0$. The Cauchy principal value (CPV) of the integral from $-\infty$ to ∞ of a function is defined by $\text{CPV}\int_{-\infty}^{\infty} f(t)\, dt = \lim_{A\to\infty} \int_{-A}^{A} f(t)\, dt$ when this limit exists. (**The reader may wish to show that for functions such that $\int_{-\infty}^{\infty} f(t)\, dt$ is absolutely convergent, the CPV gives the same value of the integral.)

If $\int_{-\infty}^{\infty} f$ is absolutely convergent, we call f **integrable** (over the interval in question, in this case the entire real line). If the integral converges with respect to some procedure but f is not integrable, it is called **conditionally convergent with respect to the procedure in question.** A description of the functions which are conditionally convergent with respect to a procedure will vary with the particular procedure. This contrasts with the simple complete description we were able to give of a conditionally convergent series.*

Example 3. The exponential distribution. Let $f(t) = ce^{-ct}$, $t > 0$, and $f(t) = 0$ otherwise, where $c > 0$. Then the kth moment of f is given by $\int_0^{\infty} t^k f(t)\, dt = c\int_0^{\infty} t^k e^{-ct}\, dt = c^{-k}\int_0^{\infty} u^k e^{-u}\, du$, where $u = ct$. Integrating by parts, we have $\int_0^{\infty} u^k e^{-u}\, du = -u^k e^{-u}\Big|_0^{\infty} + k\int_0^{\infty} u^{k-1} e^{-u}\, du = 0 + k\int_0^{\infty} u^{k-1} e^{-u}\, du$. We may use L'Hospital's rule to evaluate the first expression as zero at ∞. Continuing, we have (by induction) $\int_0^{\infty} u^k e^{-u}\, du = k!$. This is an important formula. The **factorial function** F is defined by $F(t) = \int_0^{\infty} u^t e^{-u}\, du$ for all real $t > -1$ (Exercise 10). The famous gamma function G is defined by $G(t) = F(t - 1)$ for $t > 0$. Thus the kth moment of the exponential distribution is defined for all k and is $k!/c^k$.

* One can show in advanced courses, that, analogous to the fact that a conditionally convergent series can be rearranged to converge to any value, a procedure can be formulated when I^+ and I^- are both infinite, such that the integral via this procedure gives any specified value. The idea is to choose a suitable expanding sequence of Borel sets E_n, where each E_n is a finite union of disjoint intervals, and let the procedure be $\lim_{n\to\infty} \int_{E_n} f(t)\, dt$. This freedom to arbitrarily choose the expanding sequence of Borel sets is the analog of our freedom to arbitrarily rearrange the conditionally convergent series.

EXERCISES

2.1 The sequences of ones and zeros which are ultimately 1 in each place are a countable set.

2.2 A countable union of null sets is a null set.

2.3 A piecewise continuous density function always assigns probability zero to a countable set.

2.4 Show that the Cauchy distribution is the appropriate description for the physical situations in Example 2.

2.5 Let $f(t) = 1/t$, $|t| > 1$, and $f(t) = 0$ otherwise. Find $\int_{-\infty}^{\infty} f(t)\,dt$ under the following procedures.

(*a*) $\lim_{A\to\infty} \int_{-A}^{A^2} f(t)\,dt$

(*b*) $\lim_{A\to\infty} \int_{-A}^{A+1} f(t)\,dt$

(*c*) $\lim_{A\to\infty} \int_{-A}^{Ae^2} f(t)\,dt$

(*d*) If C is an arbitrary real number, give a procedure such that $\int_{-\infty}^{\infty} f(t)\,dt = C$.

2.6 (*a*) A continuous random variable is not constant.

(*b*) The rth absolute moment of a continuous random variable, if defined, is not zero.

2.7 The second moment I_c of a continuous density function about the point c is given by the formula $I_c = V(f) + (c - E(f))^2$. Thus the second moment is strictly smaller about $E(f)$ than about any other point.

2.8 State and prove the continuous analogues of the following.

(*a*) Exercise 3.3.7.

(*b*) Exercise 3.3.8.

(*c*) Theorem 3.4.2.

(*d*) Theorem 3.4.3.

(*e*) Chebychev's inequality (Theorem 3.5.1). It is not possible for equality to hold.

(*f*) Exercise 3.5.9 (*a*).

2.9 Prove that the rth absolute moment of the Cauchy distribution exists if $0 \le r < 1$.

2.10 (*a*) Prove that the integral defining the factorial function in Example 3 converges absolutely for each fixed $t > -1$ but not for any $t \le -1$.

(*b*) Show that the rth absolute moment of the exponential distribution is defined for all $r > 0$ and given by $F(r)/c^r$.

(*c*) Show that the kth moment about the mean of the exponential distribution is $c^{-k} \sum_{j=0}^{k} (-1)^{k-j}(k)_j$, where $(k)_0 \equiv 1$.

2.11 Find the rth absolute moment of the uniform distribution about a point.

2.12 (a) A polynomial (on $(-\infty, \infty)$) is not a density function.

(*b*) When is a **rational function** (the quotient of two polynomials) a density function?

2.13 Let $f(t) = t$, $0 < t \le 1$; $f(t) = 2 - t$, $1 < t < 2$; $f(t) = 0$ otherwise. Find the kth moments of f about the mean.

2.14 Let the density function f for the continuous random variable X be defined by $f(t) = 18/11t^{10}$ when $-\infty < t \le -1$; $f(t) = -18t/11$, $-1 \le t \le 0$; $f(t) = 0$ when $t > 0$.

(*a*) Verify that $\int_{-\infty}^{\infty} f(t)\,dt = 1$.

(*b*) Graph f.

(*c*) Find $E(X)$ and $s(X)$.

(*d*) Compute the distribution function F and graph it.

3. THE NORMAL APPROXIMATION TO THE BINOMIAL PROBABILITY LAW

Our extension of probability theory to probability laws described by piecewise continuous density functions is partially motivated by the fact that many complex discrete situations are handled more simply and completely via continuous approximations. One of the earliest and best known instances of this is the approximation of the discrete binomial density by the continuous normal density. Just as the binomial probability law is the most important for finite probability, the most important continuous probability law is the normal.

To see how a piecewise continuous approximation to the binomial distribution might be discovered, consider the situation in Figure 1. The binomial distribution for $p = q = \frac{1}{2}$ and $n = 4$ is approximated by a **step function**, that is, a finite linear combination $\Sigma\, a_i C_{I_i}$ of characteristic functions of intervals. The step function used in Figure 1 has height $f_n = b(n,k)p^k q^{n-k}$ on the interval $(k - \frac{1}{2}, k + \frac{1}{2})$. It is zero for $t < -\frac{1}{2}$ and for $t > n + \frac{1}{2}$. The value is immaterial on the finite point set $-\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \cdots, n + \frac{1}{2}$ (Exercise 2.3). Then the area under the function on the interval $(k - \frac{1}{2}, k + \frac{1}{2})$ equals the value of the binomial distribution at k. Thus

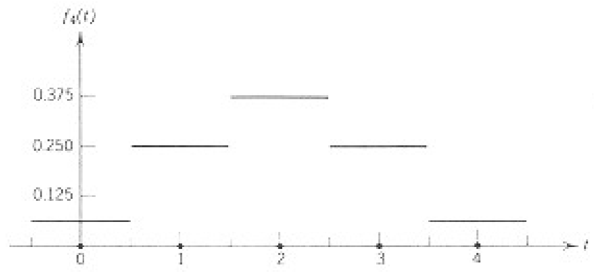


Figure 1. A piecewise continuous approximation f_4 to the binomial distribution for $p = q = \frac{1}{2}$, $n = 4$.

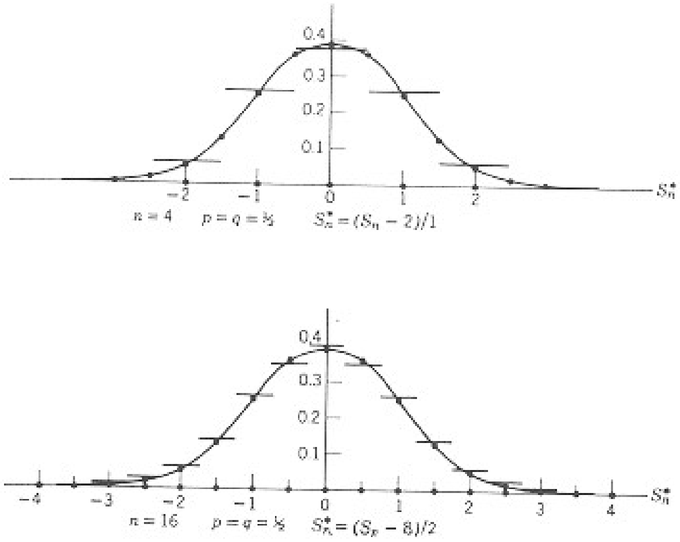


Figure 2. Approximating step functions for the density of a standardized binomially distributed random variable S_n^*. In each case the continuous curve $f(t) = e^{-t^2/2}(2\pi)^{-1/2}$ is also drawn in.

if S_n is an (n,p)-binomially distributed random variable, $P(S_n = k) = \int_{k-1/2}^{k+1/2} f_n(t)\,dt$. If E is any subset of real numbers, $P(S_n(s) \in E)$ is a suitable finite sum of such integrals. For example, $P(S_n \le \pi) = \int_{-1/2}^{1/2} f_n(t)\,dt$.

As n increases, $b(n,k)p^kq^{n-k}$ tends to zero and these piecewise continuous approximations spread out and become lower in height. To compare the approximations for various n, it is useful to study instead the standardized random variable $S_n^* = (S_n - np)/\sqrt{npq}$. Since the S_n^* all have mean zero and variance 1, we might expect their graphs to have somewhat the same scale. Figure 2 shows us what happens for the special case $p = q = \frac{1}{2}$, and $n = 4$ and 16. The heavy dots on the S_n^* axis indicate in each case the possible values of S_n^*. The approximating step functions are defined as follows. The midpoint of each step occurs at $(k - np)/\sqrt{npq}$, $k = 0, \cdots, n$. These are the possible values of S_n^*. Each step has length $1/\sqrt{npq}$, the spacing of the values of S_n^*. Thus the function is constant on the intervals $((k - \frac{1}{2} - np)/\sqrt{npq}, (k + \frac{1}{2} - np)/\sqrt{npq})$. The height of the function is chosen so that the area under a step is equal to the probability $b(n,k)p^kq^{n-k}$

that S_n^* will attain the value at the midpoint of the step. Since the step length is only $1/\sqrt{npq}$, we must multiply this probability by $\sqrt{npq}$ to get the height $\sqrt{npq}\, b(n,k)p^kq^{n-k}$ of the step.

As n increases, the step functions such as those in Figure 2 consist of steps on smaller and smaller intervals since the length of a step is $1/\sqrt{npq}$ and this tends to zero. It also turns out that the largest gap or jump from step to step tends to zero. Thus the step function looks more and more like a smooth curve. The smooth curve that the step functions tend to can be shown to be $f(t) = e^{-t^2/2}(2\pi)^{-1/2}$.

Definition 1. The normal probability density function f with parameters s and m is defined by $f(t) = e^{-(t-m)^2/2s^2}/s\sqrt{2\pi}$.

It is clear from the construction of the step functions f_n that, for each n, $P(j^* \le S_n^* \le k^*) = \int_{(j-1/2)^*}^{(k+1/2)^*} f_n(t)\,dt$, where we have used the abbreviation $r^* = (r - np)/\sqrt{npq}$, where r is any real number. If the step functions tend to $e^{-t^2/2}/\sqrt{2\pi}$ as n increases, it is reasonable to suppose that $P(j^* \le S_n^* \le k^*)$ is approximated with increasing accuracy by $(2\pi)^{-1/2}\int_{(j-1/2)^*}^{(k+1/2)^*} e^{-t^2/2}\,dt$. The following theorem to this effect was stated by de Moivre in 1733 for the case $p = 1/2$ and proved by Laplace in 1812 for all p.

Theorem 1. If S_n is an (n,p)-binomially distributed random variable and j and k are integers with $0 \le j \le k \le n$, we have

$$P(j \le S_n \le k) = \sum_{i=j}^{k} b(n,i)p^iq^{n-i} \doteq (2\pi)^{-1/2}\int_{(j-1/2)^*}^{(k+1/2)^*} e^{-t^2/2}\,dt$$

A proof of the theorem using only the techniques of a first course in calculus may be found in Neyman, pp. 234-242. It is quite lengthy so we shall omit it here. However it is very instructive as an exercise in technique. A shorter proof by a different method is given in more advanced courses.

Tables of the function $f(t) = e^{-t^2/2}(2\pi)^{-1/2}$ are widely available. Tables of the distribution function $F(t) = (2\pi)^{-1/2}\int_{-\infty}^{t} e^{-t^2/2}\,dt$ are also widely available. These are generally more useful, because of the relationship $P(j^* \le S_n \le k^*) = P(S_n \le k^*) - P(S_n < j^*) = F(k^*) - F(j^*)$. Some values of f and F appear below for reference in Tables 1 and 2. Figures 3 and 4 illustrate these data.†

Note that the change of variable $t = u^* = (u - m)/s$, where $s = \sqrt{npq}$ is the standard deviation and $m = np$ is the mean, in the integral in Theorem 1 yields

$$P(j \le S_n \le k) \doteq \frac{1}{s\sqrt{2\pi}} \int_{j-1/2}^{k+1/2} exp[-(u-m)^2/2s^2]\, du.$$

This says that an (n,p)-binomial distribution is approximately normally distributed with mean m and standard deviation s. The reason for using the form given in Theorem 1 is that the same normal distribution always appears. Thus a single table of values may be used, rather than a set of tables for various values of m and s.

A discussion of the error in the normal approximation is beyond the scope of this brief exposition. However Figure 2 gives us some feeling for it. Notice from the figure that the values of f are very close to the heights of the steps at the various values of k^*, even for such small n as 16 and 4. Furthermore, the graph of f is fairly straight over short intervals so the approximately trapezoidal area under the part of the curve over an interval $((k - \frac{1}{2})^*, (k + \frac{1}{2})^*)$ is very close to that of a rectangle with height $f(k^*)$ and this interval as a base.

The usefulness of the normal approximation will be illustrated with some examples.

† Table 1 and Figures 3 and 4 are taken from Feller, pages 165 and 167. Acknowledgment is given to Feller and Wiley for permission. Table 2 is taken from Parzen, page 441. Acknowledgment is given to Parzen and Wiley for permission.

Table 1. The normal distribution

t	$f(t)$	$F(t)$
0.0	0.398 942	0.500 000
0.1	.396 952	.539 828
0.2	.391 043	.579 260
0.3	.381 388	.617 911
0.4	.368 270	.655 422
0.5	.352 065	.691 462
0.6	.333 225	.725 747
0.7	.312 254	.758 036
0.8	.289 692	.788 145
0.9	.266 085	.815 940
1.0	.241 971	.841 345
1.1	.217 852	.864 334
1.2	.194 186	.884 930
1.3	.171 369	.903 200
1.4	.149 727	.919 243
1.5	.129 518	.933 193
1.6	.110 921	.945 201
1.7	.094 049	.955 435
1.8	.078 950	.964 070
1.9	.065 616	.971 283
2.0	.053 991	.977 250
2.1	.043 984	.982 136
2.2	.035 475	.986 097
2.3	.028 327	.989 276
2.4	.022 395	.991 802
2.5	.017 528	.993 790
2.6	.013 583	.995 339
2.7	.010 421	.996 533
2.8	.007 915	.997 445
2.9	.005 953	.998 134
3.0	.004 432	.998 650
3.1	.003 267	.999 032
3.2	.002 384	.999 313
3.3	.001 723	.999 517
3.4	.001 232	.999 663
3.5	.000 873	.999 767
3.6	.000 612	.999 841
3.7	.000 425	.999 892
3.8	.000 292	.999 928
3.9	.000 199	.999 952
4.0	.000 134	.999 968
4.1	.000 089	.999 979
4.2	.000 059	.999 987
4.3	.000 039	.999 991
4.4	.000 025	.999 995
4.5	.000 016	.999 997

Table 2. Area under the normal density function

$$\text{A table of } F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

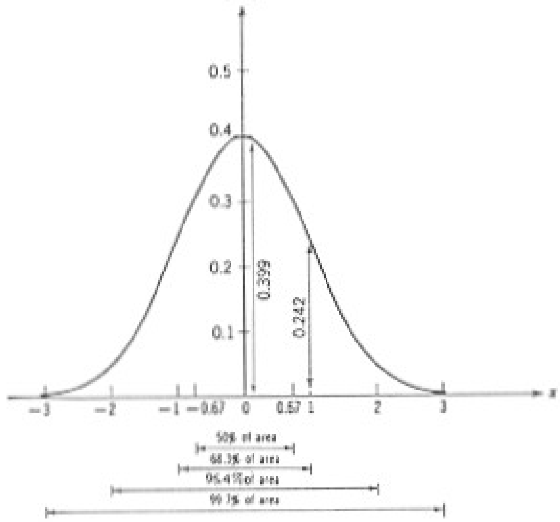


Fig. 3

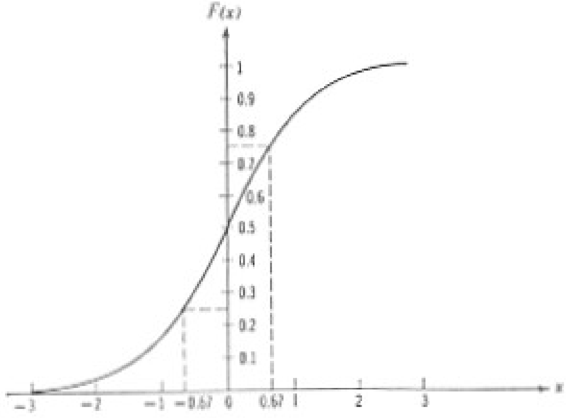


Fig. 4

Example 1. A true coin is tossed 10^4 times. What is the probability that the number of heads lies within 1% of the expected number?

The expected number is 5,000 and 1% of this is 50. Now $s = 50$ so

$$(j - \tfrac{1}{2})^* = -50.5/50 = -1.01 \qquad \text{and} \quad (k - \tfrac{1}{2})^* = 1.01$$

Thus

$$P(j \le S_n \le k) = F(1.01) - F(-1.01)$$

Table 2 only gives $F(t)$ for $t \ge 0$. However the f curve is symmetric around 0, that is,

$$f(t) = f(-t), \qquad \text{so} \quad \tfrac{1}{2} - F(-t) = F(t) - \tfrac{1}{2}, \qquad \text{or} \quad F(-t) = 1 - F(t)$$

Thus

$$F(1.01) - F(-1.01) = 2F(1.01) - 1$$

and we have the desired probability as 68.76%. If we had used the cruder limits j^* and k^* instead, we would have obtained 68.26%. If we use Chebychev's inequality, we obtain $P(|S_n/n - \tfrac{1}{2}| > 1/200) < 1$. But this follows anyhow from the fact that $P(|S_n/n - \tfrac{1}{2}| = 0) > 0$, therefore it is worthless information in this instance.

Example 2. Alois Szabo, in *The Pitfalls of Gambling and How to Avoid Them* asserts (page 32) that he has produced a profit for the last 30 years at European roulette. With a fair wheel and independent trials, a sequence of bets on the even chances (such as red or black, odd or even) is Bernoulli trials with $p = \frac{18}{37}$ and $q = \frac{19}{37}$.

Assume that a person makes 500 bets a day, 250 days a year, for 30 years. This is approximately one bet a minute, 40 hours a week, with two weeks' vacation each year. This is quite leisurely. The game is probably considerably faster than one play a minute. If these bets were on the even chances in European roulette, and the same size, what is the probability that a player would have broken even or better at the end of the 30 years?

We have $n = 3.75 \times 10^6$, $np = 1.85 \times 10^6$, and $\sqrt{npq} \doteq \sqrt{n}/2 = 968$. We seek $P(S_n \ge n/2)$. This is equivalent to

$$P\left(S_n^* \ge \frac{n/2 - np}{\sqrt{npq}}\right) \doteq P(S_n^* \ge 26.6) = 1 - F(26.6)$$

This probability is so inconceivably small that it will be amusing to

estimate it. We use the following estimate of $1 - F(t)$ from Feller (page 166):

For all

$$t > 0,\ e^{-t^2/2}(2\pi)^{-1/2}\left(\frac{1}{t} - \frac{1}{t^3}\right) < 1 - F(t) < e^{-t^2/2}(2\pi)^{-1/2}\left(\frac{1}{t}\right)$$

From this it follows that $1 - F(26.6) \doteq 10^{-156}$.

Example 3. Intelligence quotients. Psychologists tested early the abilities of children at certain age levels to reason qualitatively and quantitatively, use language, solve problems, and so forth. Roughly speaking, the overall abilities of a given child with chronological age (C.A.) A could be equated to the median overall abilities of children of a certain age. This age is called the given child's mental age (M.A.) M. The child's intelligence quotient (IQ) is defined as 100 M/A. Thus the median IQ is 100. The brighter have IQs above 100 and the duller have IQs below 100. **Intelligence** is defined (operationally) as "whatever the tests test." They are designed to test the individual's "intrinsic" mental abilities, discounting obvious educational and cultural advantages as much as is possible. For example, the test problems (other than vocabulary) should themselves be stated in language which is understandable to the widest variety of testees. (Some tests are designed to be given without the use of any language!)

It is found that the intelligence quotient remains surprisingly constant from late childhood on. Even extreme cultural advantages or disadvantages have rapidly dwindling effects as the individual matures. This comparative constancy is one of the principal values of the concept.

It is observed that the M.A.s at a certain child C.A. level are normally distributed about this C.A. as mean and with a certain standard deviation, proportional to the C.A. Thus children's IQs are normally distributed about 100 with a certain standard deviation that appears to be approximately constant.

After a C.A. of about 16, the increase in intelligence is generally small. The average adult mental age is taken to be 16.

These facts are used to extrapolate and assign adult IQs. The standard deviation appears to be about 16 IQ points for children. Scores on adult tests ("raw scores") are then assigned IQ values so that they will closely fit a normal distribution with mean 100 and standard deviation 16. Thus we may use Tables 1 and 2 or Figures 3 and 4 to describe the distribution of IQs.

What proportion of the population have IQs between 84 and 116? Since the standard deviation is 16, this is the same as the proportion lying within one standard deviation of the mean. From Figure 3, we see that this is 68.3%.

From Table 1, we see that the percentage above 132 (two standard deviations above the mean) is 2.2750%. Only 0.2350% of the people have IQs of 148 or more. Only 3 in a million have IQs of 172 or more.

It is an interesting exercise in technique, as well as useful for many applications, to compute the kth moments about the mean of the normal distribution with parameters m and s. Since

$$\frac{(2\pi)^{-1/2}}{s}\int_{-\infty}^{\infty}(t-m)^k e^{-(t-m)^2/2s^2}\,dt = (2\pi)^{-1/2}/s\int_{-\infty}^{\infty}u^k e^{-u^2/2s^2}du$$

where $t - m = u$, we may assume that $m = 0$. If we let $u/s = w$, the latter expression becomes $s^k(2\pi)^{-1/2}\int_{-\infty}^{\infty}w^k e^{-w^2/2}\,dw$ so the problem reduces to finding this last integral, I.

First, we show that the kth moment is defined for all k. Observe that $\lim_{|w|\to\infty} w^n/e^{w^2/2} = 0$ for all n. To see this, note that $e^{w^2/2} = 1 + w^2/2 + (w^2/2)^2/2! + \cdots$ is a "polynomial of infinite order." No matter what value n has, there are terms of higher degree in this expansion. Thus $e^{w^2/2} \geq c_{2m}w^{2m}$ for all m, where c_{2m} is not zero. Therefore $|w|^n/e^{w^2/2} \leq |w|^n/c_{2m}w^{2m}$ for all m and n and if $2m > n$, the latter expression, and so the former tends to zero as $|w|$ increases. Thus there is an $M > 0$ such that for $|w| \geq M$, $|w|^{k+2}e^{-w^2/2} \leq 1$, or $|w|^k e^{-w^2/2} \leq 1/w^2$. Therefore we have

$$\int_{-\infty}^{\infty}|w|^k e^{-w^2/2}\,dw \leq \int_{|w|\geq M}\frac{dw}{w^2} + \int_{|w|<M}|w|^k e^{-w^2/2}\,dw \leq \frac{2}{M} + \int_{|w|<M}M^k\,dw = \frac{2}{M} + 2M^{k+1} < \infty$$

The same proof shows that the rth absolute moment is defined for all non-negative r.

The function $f(w) = w^k e^{-w^2/2}$ is an **odd function** whenever k is odd, that is, $f(w) = -f(-w)$ for all w. If f is odd, $\int_{-A}^{A} f(w)\,dw$ is easily shown to be zero for all A. Thus the kth moments about the mean are zero for odd k.

To compute the zeroth moment (and, incidentally, verify that the normal density is really a probability density), we use a famous device. Making a suggestive choice of dummy variables, we have

$$I^2 = \int_{-\infty}^{\infty}e^{-x^2/2}\,dx\int_{-\infty}^{\infty}e^{-y^2/2}\,dy = \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty}e^{-(x^2+y^2)/2}\,dx\,dy = \int_{\theta=0}^{2\pi}\int_{r=0}^{\infty}e^{-r^2/2}r\,dr\,d\theta = 2\pi$$

Thus $I = (2\pi)^{1/2}$ and the zeroth moment is 1.

To compute the higher even moments, suppose $k = 2n$. We wish to evaluate $I_{2n} = \int_{-\infty}^{\infty} w^{2n}e^{-w^2/2}\,dw$. We can reduce the exponent on w^{2n} by an integration by parts. Let $e^{-w^2/2}w\,dw = dV$ and $w^{2n-1} = U$. Then $dU = (2n-1)w^{2n-2}$ and $V = -e^{-w^2/2}$. Therefore,

$$I_{2n} = -w^{2n-1}e^{-w^2/2}\Big|_{-\infty}^{\infty} + (2n-1)\int_{-\infty}^{\infty}w^{2n-2}e^{-w^2/2}\,dw$$

The first expression is zero by our earlier observations. Therefore $I_{2n} = (2n-1)I_{2n-2}$. An induction yields $I_{2n} = (2n-1)(2n-3)\times\cdots\times 3\times 1(2\pi)^{1/2}$.

The final result for the $2n$th even moment about the mean is $(2n-1)(2n-3)\times\cdots\times 3\times 1s^{2n}$. Some authors write $(2n-1)!!s^{2n}$ where $(2n-1)!! = (2n-1)(2n-3)\times\cdots\times 3\times 1$.

The normal approximation to the binomial distribution is a particular example of one of the most important facts in probability, the so-called Central Limit Theorem. It says that under suitable conditions the probability law of the sum $S_n = X_1 + \cdots + X_n$ of the random variables $X_1, X_2, \cdots$ is approximated more and more accurately by a normal probability law as n increases. To illustrate, we state without proof the simplest such generalization of Theorem 1.

Theorem 2. Let $X_1, X_2, \cdots$ be independent identically distributed random variables with mean m and standard deviation s. If $S_n = X_1 + \cdots + X_n$, then

$$P(a \leq S_n \leq b) = P(a^* \leq S_n^* \leq b^*) \doteq (2\pi)^{-1/2}\int_{a^*}^{b^*}e^{-t^2/2}\,dt$$

where $S_n^* = (S_n - nm)/s\sqrt{n}$ is a standardized random variable and $r^* = (r - nm)/s\sqrt{n}$ for any real number r. Note that $E(S_n) = nm$ and $s(S_n) = s\sqrt{n}$.

Extending the central limit theorem to very wide classes of situations is of great importance to probability and to statistics. This was accomplished only in this century.

As a consequence of the Central Limit Theorem, a very large number of random variables of diverse origin satisfy a normal probability law. A further example of current interest follows.

Example 4. Academic research on the stock market shows that the change in the price V of a stock can be approximately described by the normal distribution in the following way. If V_0 is the price of the stock at time $t = 0$, the price V_t at a future time $t > 0$ is described by the fact that

$\log V_t$ is normally distributed about $\log V_0$:

$$P(\log V_t \le x) = \frac{1}{s_t\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\log V_0)^2/2s_t^2}\, dx$$
$$= \frac{1}{\sqrt{2\pi a t}} \int_{-\infty}^{x} e^{-(x-\log V_0)^2/2at}\, dx,$$

where $a > 0$ is a constant of the stock known as its volatility. Note that $s_t = (at)^{1/2}$ is the standard deviation of the distribution and $\log V_0$ is the mean. (A random variable X such that $\log X$ is normally distributed is said to be **lognormally distributed**.)

The same distribution was used by Einstein to explain the ceaseless irregular Brownian motion of tiny particles suspended in a fluid (liquid or gas). There, $\log V_t$ is replaced by D_t, the displacement of the particle after time t. The constant a is called the diffusion constant. It depends on the fluid, the particle size, the temperature, and other factors.

EXERCISES

3.1 If in Example 2 the bets were all placed on single numbers, what is the probability that the player would have broken even or better at the end of 30 years?

3.2 A certain television program is designed for a mental age of 12. What percent of adults will be bright enough to watch it? What fraction of the potential adult audience would be "lost" if the program were designed for a mental age of 16? State your assumptions clearly.

Does this perhaps suggest the reason for the average quality of current television programs? Would it be even more advantageous for sponsors to increase the potential audience by further lowering the mental age for which the adult program is designed?

Do you think that sponsors who want to maximize their gross sales are likely to voluntarily raise the level for which adult television programs are designed?

3.3 (*a*) Using the estimate of $1 - F(t)$ in Example 2, show that the probability of a person having an IQ of 200 or more is $(2.16 \pm 0.03) \times 10^{-10}$. If this figure were correct, the expected number of people with IQs of at least 200, among the approximately 3×10^9 presently on earth, would be about 0.6. This is in contrast to the several such people supposed to be in the U.S. alone. Assuming that there are several people in the U.S. alone with IQs of 200 or more, the normal distribution must considerably underestimate the probability of very high IQs. (If 10^8 people in the U.S. have been tested—probably an over estimate—then the expected number is 0.02.)

(*b*) The California short-form test of mental maturity-advanced *S*-form, Grades 9 to adult (1943 revision) has 130 questions. There are 20 questions with 2 choices, 15 with 3 choices, and 95 with 4 choices. A computing machine takes the test and answers each question by making a random choice, that is,

each alternative is chosen with the same probability. What is the expected number of correct answers for the machine? This score gives the machine a mental age of 13.0, which is an adult IQ of 81. A child of 6 who is told simply to give some answer to every question, and does so, is likely to have an IQ score above 200.

Should the ability to make random marks on an answer sheet be so highly valued? Is this test well designed? Suggest a simple change in the test scoring procedure and assignment of IQs to raw scores which will completely eliminate this flaw.

3.4 Let X be an (m,s)-normally distributed random variable.

(*a*) Find $E(X^3)$.

(*b*) Find $E((X - c)^2)$.

(*c*) Find $E(X^k)$.

3.5 A certain university wishes to admit 1000 students to its next freshman class. There are 4000 applicants and experience shows that the probability that any student will accept is 0.5, and that acceptance or rejection by the various students can be assumed to be independent of each other. How many offers can the university make, and still have the probability 1% or less that more than 1000 students will accept?

Suppose instead that the university were to make 1000 offers, and then only make additional offers as rejections came in, so that there was never any risk of having more than 1000 acceptances. Other schools with bolder and more imaginative policies would make offers to the better students sooner, and end up with a higher quality selection.

3.6 *A computer study of the error in the normal approximation.* For various n, have a computer find the maximum value d_n of the magnitude of the difference between the individual terms of the binomial distribution and the corresponding estimate from the normal approximation. Plot d_n versus n.

3.7 The speed s of a gas molecule is distributed according to the Maxwell Boltzmann law $f(t) = 4\pi^{-1/2}a^{3/2}t^2e^{-at^2}$ when $t \ge 0$ and $f(t) = 0$ when $t < 0$. The constant $a = m/2kT$ where m is the mass of the molecule, T is the absolute temperature, and k is Boltzmann's constant (a universal constant, that is, independent of the physical context). Compute the kth moment about zero of this distribution.

3.8 Consider Bernoulli trials with probability $\frac{1}{2}$ of success. Let $X_n = 0$ if the nth trial results in failure and let $X_n = 2^{2n}$ if the nth trial results in success. Let $S_n = X_1 + \cdots + X_n$.

(*a*) Compute $E(S_n)$ and $s(S_n)$.

(*b*) Does S_n obey Theorem 2?

*4. PROBABILITY AND MEASURE THEORY

We end our brief introduction to probability theory by pointing out that our two great themes, discrete probability and continuous probability, are merely facets of a more general modern approach. This utilizes the theory of measure and integration. Briefly, a probability measure is

* This section is more difficult and may be omitted.

assigned to the real line by using any function F such that (compare Exercise 3.2.13*e*):

D1 F is a real-valued function defined on the entire real line.

D2 F is monotone nondecreasing, that is, $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$.

D3 $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

D4 F is right continuous at every point.

For any interval $(-\infty,b]$, we define $P((-\infty,b]) = F(b)$. Thus $P((a,b]) = F(b) - F(a)$. It is proven in advanced courses that this probability function has a unique extension to the sigma field of Borel sets.

Conversely, if P is any probability measure defined for Borel sets, we define F by $F(x) = P((-\infty,x])$. It is readily shown that F satisfies $D1$ to $D4$ (Exercise 1). There is a 1 to 1 correspondence between probability measures defined on the Borel subsets of the real line and such functions. These functions are the **distribution functions** of the general theory. We shall see that distribution functions, rather than density functions, are primary in the general theory.

The connection between these distribution functions and the ones we have studied will be partially clarified with the aid of the following theorem.

Theorem 1. If F is a distribution function (that is, satisfies $D1$ to $D4$), then

$$F = a_cF_c + a_dF_d$$

where

$$a_c \geq 0,\ a_d \geq 0,\ a_c + a_d = 1,\ F_c$$

is a distribution function that is everywhere continuous, and F_d is a discrete distribution function (that is, $F_d(t) = \sum_{x_i \leq t} f(x_i)$ where f is a discrete density function).

Proof. Note first that at each point t_0, $\lim_{t\to t_0} F(t) \equiv F(t_0^-)$ and $\lim_{t\to t_0^+} F(t) \equiv F(t_0^+)$ both exist, because of $D2$ and the fact, $D3$, that F is bounded above and below. Also $F(t_0^-) \leq F(t_0^+)$ with equality if and only if F is continuous at t_0. Since F is right continuous, $F(t_0) = F(t_0^+)$ at every point. Therefore F is discontinuous at t_0 if and only if $F(t_0^-) < F(t_0)$. The difference $F(t_0) - F(t_0^-)$ is called the **jump** $j(t_0)$ at t_0.

There are only a countable number of such points of discontinuity. To see this, let n be a positive integer. Then there are at most n points such that the jump is $1/n$ or more. If instead there were more than n such points for some n, choose $n + 1$ of them: $t_1 < \cdots < t_{n+1}$. Then $j(t_1) \leq F(t_1)$, $j(t_2) \leq F(t_2) - F(t_1), \cdots, j(t_{n+1}) \leq F(t_{n+1}) - F(t_n)$. Adding, we have $j(t_1) + \cdots + j(t_{n+1}) \leq F(t_{n+1})$. This gives $F(t_{n+1}) \geq (n + 1)/n > 1$, a contradiction. Now we can count the discontinuities: list the 1 or less points with jumps which are at least 1, then the two or less points with jumps which are at least $1/2, \cdots$, then the n or less points with jumps which are at least $1/n$, etc.

If there are no jumps, then set $a_c = 1$ and $F_c = F$. If there are one or more jumps $j(t_1), j(t_2), \cdots$, at $t_1, t_2, \cdots$, define $\sum_{t_i \leq t} j(t_i) \equiv G_d(t) > 0$. Then let $a_dF_d = G_d$ define F_d, where a_d is determined by noting that F_d satisfies $D3$. It can now be readily verified that F_d is a discrete distribution function (Exercise 2). (If there are no jumps, let $a_d = 0$.)

Now let $a_c = 1 - a_d$ and define F_c by $a_cF_c = F - a_dF_d$. It is readily verified that F_c is a distribution function and that it is everywhere continuous (a_dF_d subtracts out the jumps from F without disturbing the function elsewhere). See Exercise 3.

We have decomposed F into a linear combination of a discrete density function F_d and a continuous density function F_c. The function F_d is the discrete density function introduced in Definition 3.2.4. It is precisely the kind of object studied in Chapter 3.

The continuous density function F_c is a more complex object than we have studied in this Chapter. Thus we cannot merely apply our treatment of discrete and continuous probabilities to the separate parts of the decomposition of Theorem 1. We discuss briefly the difference between the general continuous F_c and those we have studied previously in this Chapter. It gives some hint of the complexities of the general treatment which have been avoided by our treatment.

We previously defined the distribution function by $F(x) = \int_{-\infty}^{x} f(t)\,dt$, where $f(t)$ was a piecewise continuous density function. It is easy to see that $F(x)$ is continuous for each x and that it is differentiable at all points of continuity of f, with $F'(x) = f(x)$. (Verify these statements as Exercise 4.) These functions do not include all the possibilities. In general, a continuous F_c can fail to be the integral of a piecewise continuous density function. This is one reason why distribution functions, rather than density functions, are the primary objects of the general theory.

The theory of measure and integration treats these more general continuous distribution functions. It unifies our separate treatments by constructing a more general theory of the integral. A class of functions, called measurable, which is much more inclusive than the piecewise continuous functions, is used. The new integral, called the Lebesgue integral, applies to most of these functions. It is an extension of the usual elementary (Riemann) theory in that it gives the same results for elementary integrations, has the same principal abstract properties (for example, linearity) as the elementary integration process, and lets us integrate many more functions. It will be treated in more advanced courses.

EXERCISES

4.1 Prove that the function F defined by $F(x) = P((-\infty, x])$ for each x satisfies $D1$ to $D4$.

4.2 Verify that $F_d = a_d G_d$ as defined in the proof of Theorem 1 is a discrete distribution function.

4.3 Verify that F_c is a continuous distribution function.

4.4 Verify that if F is defined by $F(x) = \int_{-\infty}^{x} f(t)\, dt$, where $f(t)$ is piecewise continuous, then:

(*a*) $F(x)$ is continuous for each x.

(*b*) $F(x)$ is differentiable at each point of continuity of f with $F'(x) = f(x)$.

References

Abbott, Edwin A., *Flatland*, Dover, New York.

Bell, E. T., *Men of Mathematics*, Simon and Schuster, New York, 1961 (paperback).

Birkhoff, Garrett, and MacLane, Saunders, *A Survey of Modern Algebra*, Third Edition, Macmillan, New York, 1965.

Cosmopolitan, September 1964, pages 10–11, E. Thorp, "Probability."

Courant, Richard, *Differential and Integral Calculus*, Volume I, Interscience, New York, 1937.

Feller, William, *An Introduction to Probability Theory and Its Applications*, Second Edition, John Wiley and Sons, New York, 1957.

Goldberg, Samuel, *Probability, An Introduction*, Prentice-Hall, 1960.

Handbook of Mathematical Tables, supplement to *Handbook of Chemistry and Physics*, First Edition, Chemical Rubber Publishing Company, Cleveland, Ohio, 1962.

Kemeny, Schleifer, Snell, and Thompson, *Finite Mathematics with Business Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

Kemeny, Snell, and Thompson, *Introduction to Finite Mathematics*, Prentice-Hall Englewood Cliffs, New Jersey, 1956.

More, Louis Trenchard, *Sir Isaac Newton*, Dover, New York, 1962 (paperback).

Mosteller, F., *Fifty Challenging Problems in Probability, with solutions*, Addison-Wesley, Reading, Mass., 1965.

Newman, James R., *The World of Mathematics*, Volumes I–IV, Simon and Schuster, New York, 1956 (paperback).

Ore, Oystein, *Cardano, The Gambling Scholar*, Princeton University Press, Princeton, N.J., 1953.

Parzen, Emanuel, *Modern Probability Theory and Its Applications*, John Wiley and Sons, New York, 1960.

Thorp, Edward, *Beat The Dealer*, Random House, New York, 1962.

Todhunter, *A History of The Mathematical Theory of Probability from the time of Pascal to Laplace*, Published in 1865 and reprinted in 1949 by Chelsea, New York.

Wilson, Allan, *The Casino Gambler's Guide*, Harper and Row, New York, 1965.

Solutions to Exercises

Chapter 1

2.1 (*a*) $1, \cdots, n$; yes.

(*b*) 0 lb, 1 lb, 2 lb, $\cdots$, 250 lb; no.

(*c*) The 48 numbers; yes if "well designed."

(*d*) 0, 1, 2, 3, 4 or 0, 1, 2, 3, 4, 5; no in either case.

(*e*) $HH, THH, THTHH, \cdots$; $TT, HTT, HTHTT, \cdots$; $THTHTH \cdots$; $HTHTHT \cdots$; no.

(*f*) $n, n+1, n+2, \cdots, 6n-1, 6n$; no.

(*g*) (H,H), (H,T), (T,H), (T,T); yes if coin true and tosses are "independent."

(*h*) All possible sequences of length n of the two symbols H,T. There are 2^n such sequences. See answer to (*g*).

(*i*) $0, 1, 2, \cdots$; no.

(*j*) (1,2), (1,3), (1,4), (2,3), (2,4), (3,4); yes.

2.2 $n \times (n-1) \times \cdots \times 2 \times 1$.

2.3 (*a*) If 00003 is considered a one digit number, there are 9×10^4. If it is considered a five digit number, there are 10^5, counting 00000.

(*b*) If 00003 is considered a one digit number, there are $9 \times 9 \times 8 \times 7 \times 6$. Otherwise there are $10 \times 9 \times 8 \times 7 \times 6$.

(*c*) $9 \times 8 \times 7 \times 6/10^4$ in either case.

3.1 (*a*) $b(n,2)$. (*b*) $b(n,3)$.

3.2 (*a*) Assuming that no spaces are allowed between symbols, there are $34 + 34^2 + \cdots + 34^6 = (34^7 - 1)/(34 - 1)$.

(*b*) $10 + \cdots + 10^6 = 1{,}111{,}110$.

3.3 (*a*) 9×10^6. (*b*) 9×10^9.

3.4 (*a*) To form precisely k couples, we can select the women in $b(68,k)$ ways. The first woman can be paired with a man in 83 ways, the second woman in 82 ways, etc. Thus there are $b(68,k)(83)_k = (68)_k(83)_k/k!$ possible ways to form the k couples.

(*b*) $\sum_{k=0}^{n} (68)_k(83)_k/k!$

(*c*) Replace 68 by w and 83 by m in the foregoing.

3.5 $b(20,8)/b(52,8) \doteq 0.9 \times 10^{-4}$.

3.7 Use the binomial theorem to expand $(1-1)^n$ and $(1+1)^n$, respectively.

3.8 There are 2^n equally likely outcomes. There are $b(n,k)$ ways in which the k "slots" with an outcome of heads can be selected, hence the probability for k heads is as given. The event E "an even number of heads" can occur in $e = \binom{n}{0} + \binom{n}{2} + \cdots$ ways. Adding the results of the preceding exercise, we find $2e = 2^n$. Thus $e = 2^{n-1}$ and $P(E) = 2^{n-1}/2^n = \frac{1}{2}$. By subtracting the first result of the preceding exercise from the second, we are led to the same answer for an odd number of heads.

3.9 The N hand can be selected in $b(52,13)$ ways from the complete deck. The S hand can be selected in $b(39,13)$ ways from the remaining 39 cards. The E hand can then be selected in $b(26,13)$ ways, and the W hand is determined by the remaining cards (or, can be selected from them in $b(13,13) = 1$ way). The result is $b(52,13)b(39,13)b(26,13)$ which is $52!/(13!)^4$ or approximately $5.36 \cdot 10^{28}$.

3.10 For $n = 2$, $k \geq 1$, the discussion is one of k-tuples in which there are two choices for each slot. Thus the situation is as in Exercise 8, where we counted the number of ways that n ($= k$ in **this** problem) coins could produce a specified number (the k of **that** problem) of heads.

For $n = 3$, $k = 3$, and ordered sampling with replacement, there are 27 outcomes. Each of the three outcomes xxx occurs once, each of the six outcomes (three choices of x, then two of y) xyy occurs three times (there are three locations for x) and each of the six outcomes xyz occurs once. Feller, II.5, gives a more extended discussion of unordered sampling with replacement. His "cells" can be thought of as "random" labels, obtained by sampling with replacement. Thus the n cells correspond to our population" of size n. His r balls correspond to our sample of size k. In particular, Feller illustrates with a detailed computation the case $n = 7$, $k = 7$.

3.12 The probabilities are given in the table at the top of page 131.

3.13 First consider all the hands with no two cards having the same rank. If we classify such hands by the ranks which they contain, we find that there are $\binom{13}{5}$ such classes, each containing 4^5 hands. But each of those classes consists of **two** equivalence classes, one corresponding to the case where the cards are all of the same suit (4 hands) and one for the case where the cards are not all of the same suit ($4^5 - 4$ hands). [*Note:* since an Ace is counted either high or low then there are 10 of the classes which

Ranking order of the hands	Number of ways hand can be drawn	Probability (in %) of drawing hand
Straight flush	40	0.00154
Four of a kind	624	0.023
Full house	3,744	0.144
Flush	5,108	0.197
Straight	10,200	0.392
Three of a kind	54,912	2.12
Two pairs	123,552	4.75
One pair	1,098,240	42.3
Nothing	1,302,540	50.2
Total number	2,598,960	

The last column is to slide-rule accuracy only; the last digit may be in error.

consist of straights or better (straight flushes). Thus we have

$$\left[\binom{13}{5} - 10\right](4^5 - 4) \quad \text{"nothing" hands,}$$

$$\left[\binom{13}{5} - 10\right]4 \quad \text{"flush" hands,}$$

$$10(4^5 - 4) \quad \text{"straight" hands,}$$

and

$$10(4) \text{ straight flush hands.}$$

There are no other possible poker hands contributing to these four types so we have completely counted them.]

Since each of our original $\binom{13}{5}$ classes consists of two equivalence classes, there are $2\binom{13}{5}$ equivalence classes of hands in which all cards have different rank. We now count the other possibilities. **Exactly one pair**: $13\binom{12}{3}$ equivalence classes with $\binom{4}{2}4^3$ hands per class. **Exactly two pairs**: $\binom{13}{2}11$ equivalence classes with $\binom{4}{2}\binom{4}{2}4$ hands per class. **Exactly three of a kind**: $13\binom{12}{2}$ equivalence classes with 4^3 hands in each class. A

full house: 13(12) classes with $4\binom{4}{2}$ hands in each class. **Four of a kind**: 13(12) classes with 1(4) hands per class. Check on the work: The total number of hands is $\binom{52}{5}$. Does

$$\binom{52}{5} = \binom{13}{5}4^5 + 13\binom{12}{3}\binom{4}{2}4^3 + \binom{13}{2}11\binom{4}{2}\binom{4}{2}4$$
$$+ 13\binom{12}{2}4^3 + 13(12)4\binom{4}{2} + 13(12)4$$

A computation shows that it does. Number of equivalence classes:

$$2\binom{13}{5} + 13\binom{12}{3} + \binom{13}{2}11 + 13\binom{12}{2} + 13(12)2 = 7462.$$

3.14 The probability of "nothing" is $\left[\binom{13}{5} - 10\right]\left(4^5 - 4\right)\Big/\binom{52}{5} =$ $(1277 \times 1020)/2{,}598{,}960 = 1{,}302{,}540/2{,}598{,}960$. Thus the "nothing" hands contain 3060 more than half the total and the median will lie among the "nothing" hands. Each of the "nothing" equivalence classes has $(4^5 - 4) =$ 1020 members so we must count down 3060/1020 or three equivalence classes, and we see to our surprise that the seemingly unlikely possibility for the location of the median holds. The four highest ranking "nothing" hands are, in order, A K Q J 9, A K Q J 8, A K Q J 7 and A K Q J 6, with the cards drawn from at least two suits. Thus the median is between A K Q J 7 and A K Q J 6. Precisely half the hands are at least as good as A K Q J 7 and precisely half are as poor as A K Q J 6.

3.15 (*a*) $b(51,12)/b(52,13) = \frac{13}{52} = \frac{1}{4}$, in agreement with the argument that the four events N, S, E, or W respectively get the Ace of spades are equally likely.

(*b*) $b(39,13)/b(52,13)$.

(*c*) $1/b(52,13)$.

(*d*) $4/b(52,13)$.

Chapter 2

1.1 (*a*) 12, 3: $B \subset A$ gives 12; $AB' = 11$, $AB = 3$, $A'B = 9$, $A'B' = \varnothing$ gives 3.

(*e*) 23, 14; rest as in (*a*).

1.2 (*a*) AB. (*b*) $A \cup B$.

1.3 $\varnothing$, {1}, {2}, {3}, {1,2}, {1,3}, {2,3}, S.

1.4 We want the least n such that 2^n is at least 10^{79}. Taking logarithms gives $n \log_{10}2 \geq 79$ or $n \geq 263$.

***1.11** $8 = 2 * 3 \neq 3 * 2 = 9$ so $*$ is not commutative. $256 = 2 * (2 * 3) \neq (2 * 2) * 3 = 64$ so $*$ is not associative. $1 = 1 * (1 + 1) \neq (1 * 1) + (1 * 1) = 2$, and $4 = (1 + 1) * 2 \neq (1 * 2) + (1 * 2) = 2$ so neither distributive law holds.

***1.12** The first two sets to be combined can be selected in $(n)_2$ ways. Their order in combining is immaterial because of commutativity. The resulting set can be combined with a third set, chosen from the remaining $n - 2$ sets, in $n - 2$ ways. Continuing, we have $(n)_2(n - 2)!$ ways of computing, or $(n!)/2$. Note that some possible ways of combining the sets have been omitted when $n \geq 4$. With four sets, for instance, the computations of the type $(A \cup B) \cup (C \cup D)$, of which there are six when we rearrange the parentheses, have been omitted. We have only counted those of the type $((A \cup B) \cup C) \cup D$.

****1.13** (*a*) A lawyer might argue as follows. (One did, to me.) Suppose neither the first man nor the second man knows. The first man must have seen at least one red hat. If the third man had a black hat, the second man, knowing this, would know he had a red hat. But he too didn't know. Therefore the third man knows that he has a red hat.

(*b*) The mathematician often solves problems by first considering an extreme special case. Let us begin with the simplest case, that of two men, two red hats, and a black hat. The three possible arrangements of hats are RR, RB, and BR, where the first letter denotes the color of the first man's hat and the second letter denotes the color of the second man's hat.

The first man will know his hat color only in the event of RB. The second man, knowing this, then also knows his hat color. For if the first man does not know, RB is eliminated as a possibility for the second man. In both the remaining combinations, the second man has a red hat.

For the original problem, the cases and the outcomes after two men have answered are:

1	2	3	
R	R	R	1 and 2 don't know
R	B	R	
B	R	R	
B	B	R	
R	R	B	2 knows by deduction
R	B	B	1 knows by observation
B	R	B	2 knows by observation

Notice that we have grouped six of the seven cases into pairs that appear to be the same to 2. When 2 eliminates RBB because 1 would have known it by observation, he can then assert that if he observes RXB, where X is unknown, then X is in fact R. This is the key to the problem. Suppose, for example, that 1 and 2 don't know. We see that 3 can now assert he has a red hat.

The general case of n men, n red hats and $n-1$ black hats proceeds in the same way. Man 1 eliminates one combination by observation, and man 2 eliminates one more combination by deduction and one more combination by observation. The kth man finds $2^0 + 2^1 + \cdots + 2^{k-2} = 2^{k-1} - 1$ combinations already eliminated. He uses these to eliminate 2^{k-1} more combinations, $2^{k-1} - 1$ of them by deduction and 1 more by observation, for a total of $2^k - 1$ combinations eliminated after the kth man answers. Thus, by the time the nth man answers, $2^n - 1$ combinations, or all possible combinations, will be eliminated.

The proof can be formalized by using mathematical induction. It may be helpful to think of B and R as 0 and 1, respectively, and the combinations as the binary numbers from 1 to $2^n - 1$.

(*c*) If the number of black and red hats both are greater than or equal to n, all 2^n combinations are possible. No deductions can be made other than the n from observation. Since $n < 2^n$ for positive n, the assertion of the problem is false in this case.

If $r > n$, $b = n - 1$, precisely the same combinations appear and the solution is unchanged. If $r \geq n$, $b < n - 1$, some but not all of the combinations for the case $r = n$, $b = n - 1$ appear and the solution applies in the same way to the remaining cases. The same reasoning holds when the roles of b and r are interchanged.

2.1 There are three such non-empty functions, each of them constant, corresponding to the three non-empty subsets of X which are possible domains: (1,1); (2,1); (1,1), (2,1). Adding the empty function gives four functions in all.

2.2 2^n.

2.3 If $D = \{1\}$, the two possible functions are (1,1) and (1,2). If $D = \{2\}$, the two possible functions are (2,1) and (2,2). If $D = \{1,2\}$, the four possible functions are (1,1), (2,1); (1,1), (2,2); (1,2), (2,1); (1,2), (2,2). Thus there are eight possible non-empty functions. All the functions are 1–1 except (1,1), (2,1) and (1,2), (2,2). Only (1,1), (2,2) and (1,2),(2,1) are onto. With the empty function there are nine functions.

2.4 A function from X to Y maps each point of X into 1, 2, ..., n or nowhere. Thus there are $n + 1$ choices for each point of X, hence $(n+1)^m$ functions.

3.3 $P(Y) = b(32,13)/b(52,13)$. An (optional) calculation shows that $P(Y) \doteq 1/1828$ so the Duke had a good bet.

3.4 The probability is $b(36,13)/b(52,13) \doteq 3.64 \times 10^{-3} \doteq 1/275$.

3.6 $P(B) = b(m,2)/b(m+n, 2)$ $P(W) = b(n,2)/b(m+n, 2)$ $P(M) = b(m,1)b(n,1)/b(m+n,2)$.

4.3 (*a*) There are n^k ways to assign birthdays to k people, $(n)_k$ of which are distinct, where we are to take $n = 365$. Thus the probability of the event E that no two people have the same birthday is $(n)_k/n^k = (1 - 1/n)(1 - 2/n) \cdots (1 - (n - k + 1)/n)$. The probability that some two people have the same birthday is $P(E') = 1 - P(E)$.

(*b*) There are $(n-1)^{k-1}$ ways to assign birthdays to the other $k - 1$ people so that none of them have your birthday. The probability is therefore $(n-1)^{k-1}/n^{k-1}$.

4.4 In Example 1.3.3, let (1,2), (1,3) and (2,3) each have probability $\frac{1}{3}$, and let (2,1), (3,1), and (3,2) each have probability 0.

4.5 If we drop $P(S) = 1$, then an example is any function $f = cP$ defined by $f(E) = cP(E)$ for all E in S, where $c \neq 1$ is a non-negative real number and P is a probability measure.

If we drop $P(E) \geq 0$, then an example can be made up as follows. If $E_1, \cdots, E_n$ are elementary events, let $f(E_i) = p_i$ where $\sum_{i=1}^{n} p_i = 1$ and at least one p_i is negative. Define $f(E)$ for an arbitrary subset E as $\Sigma f(E_i)$ with the sum taken over all $E_i \subset E$. Such an example is possible if and only if S has at least two points.

If we drop $P3$, then any function f such that $f(S) = 1$ and $f(E) \geq 0$ for each E in S is such an example.

In each of the above cases, we have given all the examples which are possible. The student may supply the proof of this fact.

5.1 $P(U_i) = 1/(n+1)$, $P(W) = \frac{1}{2}$, and $P(U_iW) = i/n(n+1)$. Independence gives $i/n = 1/2$ or $i = n/2$. Thus independence occurs only when n is even and $i = n/2$.

5.2 none; (*c*), (*d*), (*e*).

5.3 Let D, R, U, and F be the events the voter is a Democrat, Republican, unaffiliated, or favors the bond issue, respectively. We seek $P(D \mid F)$, $P(R \mid F)$, and $P(U \mid F)$. Now $P(F) = P(FD) + P(FR) + P(FU) = 0.525$. Whence $P(D \mid F) = P(DF)/P(F) = 0.4/0.525 = 0.762$, $P(R \mid F) = 0.152$, and $P(U \mid F) = 0.086$.

5.4 no; yes.

5.5 $b(39,13)/b(52,13)$; $b(32,13)/b(39,13)$; $b(24,13)/b(26,13)$; $\geq b(24,13)/b(26,13)$.

5.6 Let $S = \{1,2,3\}$. Suppose A_1, A_2, A_3 is a collection of pairwise independent subsets of S with $0 < P(A_i) < 1$. Then each A_i has 1 or 2 points. But $A_i \not\subset A_j$ for $i \neq j$ (observation (*e*) following definition 2) so each A_i has 2 points and the A_i are distinct. Further, by relabeling the points of S if necessary, it follows that we may assume $A_1 = \{2,3\}$, $A_2 = \{3,1\}$, and $A_3 = \{1,2\}$.

Setting $P(\{i\}) = p_i$, we now have the following conditions on the p_i.

Condition	Reason
(1) $p_1 + p_2 + p_3 = 1$	$P(S) = 1$
(2) $(p_2 + p_3)(p_3 + p_1) = p_3$	$P(A_1)P(A_2) = P(A_1A_2)$
(3) $(p_3 + p_1)(p_1 + p_2) = p_1$	$P(A_2)P(A_3) = P(A_2A_3)$
(4) $(p_1 + p_2)(p_2 + p_3) = p_2$	$P(A_3)P(A_1) = P(A_3A_1)$
(2') $(1 - p_1)(1 - p_2) = p_3$	(1) and (2)
(3') $(1 - p_2)(1 - p_3) = p_1$	(1) and (3)
(4') $(1 - p_3)(1 - p_1) = p_2$	(1) and (4)

Using (1), (2') to (4') simplify to $p_ip_j = 0$, $i \neq j$. But this means that at least two of the p_i are zero. From (1) it follows that the third p_i is one. But then $P(A_i) = 1$ for those A_i containing this p_i. This contradicts our assumption $P(A_i) < 1$ for all i. Therefore such an example is not possible.

The remainder of the problem now follows with the aid of the observations (*a*) to (*g*) following Definition 2.

5.7 $P(A')P(B) = (1 - P(A))P(B) = P(B) - P(A)P(B) = P(B) - P(AB) = P(A'B)$. Similarly for the other assertions.

5.8 1/11.

5.9 If we consider the samples as ordered, there are n^k equally likely outcomes. Suppose precisely j of the results of the samples are a 1. There are $(n-1)^{k-j}$ outcomes if we specify which j members of the k-tuple are 1. There are $b(k,j)$ ways to specify the j ones. Thus there are $b(k,j)(n-1)^{k-j}$ outcomes with precisely j ones. The number with at least j ones is therefore $f(j) = \sum_{i=j}^{k} b(k,i)(n-1)^{k-i}$ and the probability of at least j ones, given at least $j - 1$ ones, is $f(j)/f(j-1)$. Applying this to the preceding problem as a check, we have $f(1) = b(2,1)\cdot 5 + b(2,2) = 11$ and $f(2) = b(2,2) = 1$ so the desired probability is found to be 1/11 as before.

5.11 Goldberg, page 81.

5.12 We assume random sampling throughout. There are $b(13,5)$ flushes per suit and four suits so there are $4b(13,5)$ flushes. There are $b(52,5)$ hands so $P(A) = 4b(13,5)/b(52,5)$ or $4(13)_5/(52)_5$. Suppose now that player 1 has been dealt a flush. Player 2 can be dealt $b(8,5)$ flushes from player 1's suit and $3b(13,5)$ flushes from the other suits. His hands are dealt from the 47 remaining cards so he can receive $b(47,5)$ hands. Therefore $P(B \mid A) = (b(8,5) + 3b(13,5))/b(47,5) = ((8)_5 + 3(13)_5)/(47)_5$. Intuition suggests that $P(B \mid A) < P(A)$: since player 1 has used up one of the possible rare flushes, it should be harder for player 2 to draw one. A computation shows that $P(B \mid A) \doteq 2.55 \times 10^{-3}$ and $P(A) \doteq 1.98 \times 10^{-3}$, in defiance of the intuitive argument above.

My (nonmathematician) wife gave the following answer when I later asked her whether she would expect $P(B \mid A)$ to be greater than, equal to, or less than $P(A)$.

> I know as a bridge player, both from my own experience and the reported experience of other players, that when there is a freakish hand (one with a great excess of cards in one or more suits and a corresponding scarcity of cards in one or more other suits) at the table, the chances are increased that there will be another freakish hand as well. Therefore I would say that $P(B \mid A) > P(A)$.
>
> However, I would say that most people would argue that since one flush is gone, there aren't as many left now so the chances are smaller. I still think $P(B \mid A) > P(A)$. What is the answer, and can you explain it without using mathematics?

In the case of the flush, we observe that the number of flushes available to player 2 is about $3b(13,5)/4b(13,5)$—we can neglect $b(8,5)$—or about 75% as many as were available to player 1. But the total number of hands available to player 2 is $b(47,5)/b(52,5) \doteq 0.9^5 \doteq 0.6$ or 60% as many as were available to player 1. Thus the ratio $P(B \mid A)/P(A)$ is about 0.75/0.60 or 1.25 (more precisely, 2.55/1.98 or 1.29). Note in contrast that if eight of nine players all are dealt flushes, it is impossible for the ninth player to have a flush. But in bridge, if three of four players are dealt "flushes," the fourth player must have a "flush."

5.13 Let S be the eight-point sample space of outcomes when three coins are tossed. Let P_1 be the equiprobable measure. If H_i is the event "the ith coin falls heads," then the H_i are mutually independent with respect to (S,P_1).

There are many ways to choose P_2 so that the H_i are not pairwise independent with respect to (S,P_2). For instance, let (H,T,T), (T,H,T), and (H,H,T) each have probability 1/3. Let the other points each have probability 0. Then $P(H_1) = P(H_2) = 2/3$ and $P(H_1H_2) = 1/3 \neq P(H_1)P(H_2) = 4/9$.

6.1 Let C be the event that the expert was cheated. Let L be the event that he loses 20 units in an hour. Then $P(L \mid C) = 1$, $P(L \mid C') = 0.25$, and $P(C) = 0.1$. We wish to find $P(C \mid L)$. Considering C and C' as "causes," Bayes' formula gives

$$P(C \mid L) = \frac{P(L \mid C)P(C)}{(P(L \mid C)P(C) + P(L \mid C')P(C'))}$$

$$= \frac{0.1}{(0.1 + 0.225)} = \frac{4}{13} \doteq 0.308$$

6.3 We have $P(H_1) = 10^{-6}$ and $P(H_2) = 1 - 10^{-6} \doteq 1$. We seek $P(H_2 \mid E)$. Now $P(E \mid H_1) = 1$ and $P(E \mid H_2) = 2^{-20} \doteq 10^{-6}$. Thus, from

Bayes' theorem, we have

$$P(H_2 \mid E) = \frac{2^{-30}}{(10^{-6} + (1 - 10^{-6})2^{-30})} \doteq \frac{10^{-9}}{(10^{-6} + 10^{-9})} \doteq 10^{-3} = 0.001$$

The reader should assure himself that the approximations introduce no errors in the first significant figure.

Note that if two-headed coins were rarer, then $P(H_2 \mid E)$ would be greater. For instance, if $P(H_1) = 10^{-9}$, then $P(H_2 \mid E) \doteq 1/2$.

6.4 We note that $i^3 - (i-1)^3 = 3i^2 - 3i + 1 = 3(i)_2 + 1$. Summing from 1 to n we have $n^3 = 3\sum_{i=1}^{n}(i)_2 + \sum_{i=1}^{n} 1$. Noting that $\sum_{i=0}^{n}(i)_2 = \sum_{i=1}^{n}(i)_2$, the result follows.

6.5 By symmetry (think this through carefully), we can write down at once $P(U_i \mid B) = P(U_i \mid W)$.

There are $(n)_2$ **ordered** ways to choose two balls without replacement from n balls. There are $i(n-i)$ ways in which the first ball is white and the second is black. Likewise, there are $(n-i)i$ ways in which the first ball is black and the second is white. Thus $P(M \mid U_i) = 2i(n-i)/(n)_2$. Noting that the $P(U_i)$ are equal and cancel throughout, Bayes' rule yields

$$P(U_i \mid M) = P(M \mid U_i) \Big/ \sum_{i=0}^{n} P(M \mid U_i) = i(n-i) \Big/ \sum_{i=0}^{n} i(n-i)$$

The denominator can be evaluated if desired by the technique of Courant 1.4, Example 1. The result is $(n+1)_3/6$.

We see that the numerator is greatest when i is nearest to $n/2$ and decreases as i gets farther from $n/2$, in agreement with intuition. Note too that M is impossible when $i = 0$ or $i = n$. As a further check on the result, observe that $\sum_{i=0}^{n} P(U_i \mid W)$, $\sum_{i=0}^{n} P(U_i \mid B)$, and $\sum_{i=0}^{n} P(U_i \mid M)$ all equal 1, in agreement with Exercise 5.10.

6.6 We seek those values of p for which $P(S \mid F_s) = 0.8p/(0.8p + 0.4(1-p)) > p$. This is equivalent to $p > 0$ and $0.8 > 0.8p + 0.4 - 0.4p$, which is equivalent to $0 < p < 1$. Thus a forecast of sunny always causes the student to revise his estimate upward, except in the extreme cases $p = 0$ and $p = 1$, where he knows for sure what will happen.

When $p = 1$, $P(S \mid F_s) = 1$, the greatest value possible (by Exercise 5.10). If $p < 1$, it is readily seen from the expression above for $P(S \mid F_s)$ that it is less than 1. This is consistent with what we expect. If the student predicts sunny weather with probability 1 and the figure is accurate (as we assumed), then no matter how accurate the forecaster is, it ought to be sunny. (We assume, however, that $P(F_s) > 0$, in order to have $P(S \mid F_s)$ defined.) If both the student and the forecaster predict a sunny day, but both have some uncertainty in their prediction, then it should not be certain that the weather will be sunny, that is, $P(S \mid F_s) < 1$ when $p < 1$ is to be expected.

6.7 $(i+1)_{r+1} - (i)_{r+1} = (r+1)(i)_r$. Summing from $i = 1$ to n yields the desired relation. Comparing this with Exercise 4, we find that solving the general problem may be easier, once we have posed it, than solving a special case. This is a common occurrence in mathematics.

7.5 The following conditions must be satisfied by a and b: (1) $0 < a, b < 1$, (2) $ab = 1/3$, (3) $(1-a)(1-b) = 1/3$, (4) $(1-a)b + a(1-b) = 1/3$.

Using (2) to simplify (3) yields (3'): $a + b = 1$. Now (2) and (3') yield (4) so it contains no new information and may be discarded. Eliminating b from (2) and (3') and solving the resulting equation for a shows that no real number a satisfies the equation. Therefore the given conditions cannot be satisfied.

8.2 Feller, page 89.

8.3 We have

$$P(A_i) = \frac{b(39,13)}{b(52,13)} \doteq 1.28 \times 10^{-2}$$

$$P(A_iA_j) = \frac{b(26,13)}{b(52,13)} \doteq 1.64 \times 10^{-5} \qquad i < j$$

$$P(A_iA_jA_k) = 1/b(52,13) \doteq 1.6 \times 10^{-12} \qquad i < j < k.$$

Substituting, Theorem 1 yields $P(\bigcup_{i=1}^{4} A_i) \doteq 5.11 \times 10^{-2}$.

Chapter 3

1.1 If $p = 0$, $P(E) = 1$ if $1 \in E$ and $P(E) = 0$ if $1 \in E'$. If $p = 1$, $P(E) = 1$ if $\infty \in E$ and $P(E) = 0$ otherwise.

1.7 A true coin could be selected and tossed until a tail appears. The number of the trial on which this occurs is the number of the urn to be selected.

$$P(U_n) = \frac{1}{2^n} \quad \text{and} \quad P(W \mid U_n) = \frac{1}{n}$$

Thus

$$\sum_{n=1}^{\infty} P(U_n)P(W \mid U_n) = \sum_{n=1}^{\infty} \frac{1}{n2^n}$$

This last sum is $\log_e 2$. Using Bayes' rule, as stated in the preceding exercise, we have $P(U_n \mid W) = 1/n2^n \log_e 2$.

To evaluate the sum, note that

$$-\log(1-x) = \int_0^x \frac{dt}{(1-t)} = \int_0^x \sum_{n=1}^{\infty} t^{n-1}\, dt = \sum_{n=1}^{\infty} \int_0^x t^{n-1}\, dt = \sum_{n=1}^{\infty} \frac{x^n}{n}$$

when $|x| < 1$. The interchange of the order of summation and integration is justified because the sum is uniformly convergent for $0 \leq t \leq |x| < 1$. Setting $x = \frac{1}{2}$ gives the result.

2.1 2^Y consists of $\varnothing$ and Y only. $f^{-1}(\varnothing) = \varnothing$ and $f^{-1}(Y) = D$. Any other function from 2^Y to 2^X is not an f^{-1} for some f. For example, let g be a constant function such as that defined by $g(\varnothing) = X$ and $g(X) = X$.

2.2 The domain of f^{-1} is always 2^Y, which in this case is the collection $\varnothing$, $\{3\}$, $\{4\}$, Y. We have $f^{-1}(\varnothing) = \{x: f(x) \in \varnothing\} = \varnothing$, $f^{-1}(\{3\}) = \{x: f(x) \in \{3\}\} = \{1,2\} = X$, $f^{-1}(\{4\}) = \{x: f(x) \in \{4\}\} = \varnothing$, and $f^{-1}(Y) = \{x: f(x) \in Y\} = D = X$.

2.6 Let $R_1, \cdots, R_n$ be subsets of the real numbers. Then $X_i^{-1}(R_i)$ is a product set of the form E_i^*, $E_i \subset S_i$. We are to show that the E_i^* are independent. If $i_1 < \cdots < i_k$ is a subcollection of distinct indices, then $E_{i_1}^* \cap \cdots \cap E_{i_k}^*$ is the product set with ith term E_{i_j} if $i = i_j$ for some j, and ith term S_i otherwise. Thus

$$P(E_{i_1}^* \cap \cdots \cap E_{i_k}^*) = P_{i_1}(E_{i_1}) \cdots P_{i_k}(E_{i_k}) = P(E_{i_1}^*) \cdots P(E_{i_k}^*)$$

where the equalities follow from the definition of the cartesian product measure on product sets.

2.7 No. For instance, for $i \neq j$ we have $P(X_i = n)P(X_j = n) \neq P(X_i = n, X_j = n) = 0$.

2.8 (*a*) If $C_1, \cdots, C_n$ is a collection of constant random variables and $R_1, \cdots, R_n$ are subsets of real numbers, then $C_i^{-1}(R_i) = S$ or $\varnothing$ for each i. Such a collection of sets is readily seen to be independent (observation (*d*) following Definition 2.5.2).

(*b*) Reduces to (*d*) following Definition 2.5.2.

(*c*) Reduces to (*c*) following Definition 2.5.2.

(*d*) Follows from (*c*) and (*d*) following Definition 2.5.2.

(*e*) Use (*d*) and Example 2.5.3.

2.9 Choose a subset of distinct indices, numbered $1, \cdots, m$ for notational convenience (the proof is identical for all other choices of sets of distinct indices so we obtain full generality). If $R_1, \cdots, R_m$ are subsets of real numbers, $X_i^{-1}(R_i) = X_i^{-1} \circ f_i^{-1}(f_i(R_i)) = (f_i \circ X_i)^{-1}(f_i(R_i))$. But these latter sets are independent because the $f_i \circ X_i$ are independent random variables by hypothesis and the $f_i(R_i)$ are sets of real numbers. Hence the $X_i^{-1}(R_i)$, and thus the X_i, are independent.

2.10 Let $X_k(S) = \{x_{k,1}, \cdots, x_{k,n_k}, \cdots\}$. Let $\{R_k\}$ be a collection of one point subsets of real numbers, that is, $R_k = \{x_{k,j}\}$ for each k and some (one) j depending on k. There are at least Πn_k distinct ordered collections of real numbers that can be made in this way. For each such collection we have $P(\bigcap_{j=1}^{n} X^{-1}(x_{k,j})) = \prod_{j=1}^{n} P(X = x_{k,j}) \neq 0$; therefore each of the sets $\bigcap_{j=1}^{n} X^{-1}(x_{k,j})$ contains at least one point. These sets are also disjoint,

as follows from Exercise 2.15. Since each of these Πn_i disjoint sets contains at least one point, the collection contains at least Πn_i points.

2.18 Let (S,P_1) be as in Example 6, where (S,P_1) is a cartesian product of n spaces, each of which has the two outcomes T and H, with $P(H) = p$, $p(T) = q = 1 - p$. Now let P_2 be defined by $P_2(H, \cdots, H) = P_2(T, \cdots, T) = \frac{1}{2}$, and P_2 is zero otherwise. Then $P(X_i = 1) = \frac{1}{2}$ and for distinct indices i and j, $P(X_i = 1, X_j = 1) = \frac{1}{2}$ so $P(X_i = 1)P(X_j = 1) \neq P(X_i = 1, X_j = 1)$. The X_i are not even pairwise independent. (It is enough to show that some pair is not independent. We have found even more: no pair is independent.)

3.2 Group the positive and negative terms in "blocks" as follows. The mth block P_m of positive terms is $P_1 = \frac{1}{4}$; $P_2 = \frac{1}{6} + \frac{1}{8} > \frac{1}{4}$; $P_3 = \frac{1}{10} + \cdots + \frac{1}{16} > \frac{1}{4}$; and in general $P_m = 1/(2^m + 2) + \cdots + 1/2^{m+1} > 2^{m-1}(1/2^{m+1}) = \frac{1}{4}$. The mth block N_m of negative terms is $N_1 = -\frac{1}{3} < -\frac{1}{4}$; $N_2 = -\frac{1}{5} - \frac{1}{7} < -\frac{1}{4}$; $N_3 = -\frac{1}{9} - \cdots - \frac{1}{15} < -\frac{1}{4}$; and in general $N_m = -1/(2^m + 1) - \cdots - 1/(2^{m+1} - 1) < -\frac{1}{4}$.

The following arrangement diverges to $+\infty$:

$$P_1 + \cdots + P_8 - 1 + P_9 + \cdots + P_{16} - \tfrac{1}{3} + \cdots > 1 + 1 + \cdots.$$

The following arrangement diverges to $-\infty$:

$$N_1 + \cdots + N_8 + \tfrac{1}{2} + N_9 + \cdots + N_{16} + \tfrac{1}{4} + \cdots.$$

The following arrangement diverges because it oscillates:

$$-1 + \tfrac{1}{2} + N_1 + P_1 + N_2 + P_2 + \cdots.$$

3.8 $E(a_i C_{E_i}) = a_i P(a_i C_{E_i} = a_i) = a_i P(C_{E_i} = 1) = a_i P(E_i)$. Now sum over i.

3.9 (i) implies (ii): The expressions for $E(X^+)$ and $E(X^-)$ are subseries of the series for $E(X)$ and therefore are also absolutely convergent.

(ii) implies (iii): Theorem 1.

(iii) implies (i): The series for $E(|X|)$ is the absolute value, term by term, of the series for $E(X)$. Convergence of the $E(|X|)$ series means absolute convergence of the $E(X)$ series.

3.11 To show that $E(X + Y) = E(X) + E(Y)$, let $X(S) = \{x_1, x_2, \cdots\}$ and $Y(S) = \{y_1, y_2, \cdots\}$. Define $A_i = X^{-1}(x_i)$, $B_j = Y^{-1}(y_j)$. Then $\{A_i\}$, $\{B_j\}$, and $\{A_iB_j\}$ are finite or countably infinite partitions of S (Exercise 3.4). We have

$$\begin{aligned} E(X + Y) &= \sum_{i,j} (x_i + y_j)P(A_iB_j) = \sum_i x_i \sum_j P(A_iB_j) + \sum_j y_j \sum_i P(A_iB_j) \\ &= \sum_i x_i P\Big(\sum_j A_iB_j\Big) + \sum_j y_j P\Big(\sum_i A_iB_j\Big) \\ &= \sum_i x_i P(A_i) + \sum_j y_j P(B_j) = E(X) + E(Y) \end{aligned}$$

We have written the equalities in the order which motivates their discovery. For the proof, we read them backwards: $E(X)$ and $E(Y)$ exist and equal the given sums, $\cdots$, etc. The expression for $E(X+Y)$ is absolutely convergent because the sum of two absolutely convergent series is absolutely convergent.

3.14 (*a*) The probability that the first run begins with a success and has length precisely k is p^kq, where $k = 1, 2, \cdots$. Similarly, the probability that it begins with a failure and has length precisely k is q^kp. Thus

$$E(L) = \sum_{k>0} k(p^kq + q^kp) = pq\left(\sum_{k>0} kp^{k-1} + \sum_{k>0} kq^{k-1}\right)$$

Call the first sum h and the second sum t. Observe that $\sum_{k\geq 0} x^k = (1-x)^{-1}$ is absolutely convergent for $|x| < 1$. Thus it can be differentiated term by term, yielding $\sum_{k>0} kx^{k-1} = (1-x)^{-2}$ for $|x| < 1$. Hence $h = q^{-2}$ and $t = p^{-2}$ so $E(L) = p/q + q/p$.

The minimum is for $p = \frac{1}{2}$. When p approaches 0 or 1, $E(L)$ approaches infinity. To find the minimum without differentiation, guess (by symmetry of $g(p) = p/q + q/p$ about $p = ½$) that $g(p)$ has a unique minimum at $p = ½$. Now prove this by showing $g(p) - g(½) = (2p-1)^2/p(1-p) \geq 0$ with equality only at $p = ½$.

(*b*) Suppose p is near 1. Then the first run is a success run, with probability p. It tends to be quite long. The second run will then be a failure run and will tend to be quite short. The chance that the second run is a (comparatively long) success run depends on the first run being a failure run. This has the small probability q.

(*c*) $E(L_2) = p(pqt) + q(pqh) = q + p = 1$.

4.2 $E((X-c)^2) = E(X^2) - 2cE(X) + c^2$. The first derivative is $-2E(X) + 2c$. Thus the only possible minimum is at $c = E(X)$. The second derivative is $2 > 0$ so $E(X)$ is a minimum.

Here is a proof that does not involve differentiation

$$E((X-c)^2) = E(X^2) - 2cE(X) + c^2 \quad\text{and}\quad V(X) = E(X^2) - E(X)^2$$

Subtracting the second equation and factoring the right side of the result gives $E((X-c)^2) - V(X) = (c - E(X))^2 \geq 0$. The right side is zero if and only if $c = E(X)$, and the result follows.

4.3 One example is: $f(1) = f(-1) = \frac{1}{2}$, and f is zero otherwise. Any density function with the property $f(t) = f(-t)$ for all real t will do. Prove this.

4.6 Since Y is not a constant, $f_X(0) < 1$. Thus there is an $x_i \neq 0$ such that $f_X(x_i) > 0$. Since f_X is even, $f_X(-x_i) = f_X(x_i) = p$ where $0 < p < \frac{1}{2}$. The case $p = \frac{1}{2}$ is excluded because it would imply that Y is a constant, contrary to assumption.

Let $E = \{-x_i, x_i\}$, $F = \{x_i^{2n}\}$. Then $X^{-1}(E) = 2p$, $Y^{-1}(F) = 2p$, and $X^{-1}(E)Y^{-1}(F) = 2p$. In order for the sets to be independent, $(2p)^2 = 2p$. Since $p \neq 0$, this simplifies to $p = \frac{1}{2}$, which has been excluded. Thus X and X^{2n} are dependent whenever X^{2n} is not constant.

4.7 (*a*) $f(x) = |x|$ is a pair of rays issuing from the origin at 45° and 135°, or slopes of 1 and -1, respectively. $f(x) = |x|/n$ is a pair of such rays with slope $\pm 1/n$. $f(x) = |x-a|/n$ is the graph of $f(x) = |x|/n$ translated to the right a units.

The functions are continuous everywhere and differentiable everywhere except at the vertex (that is, where $f(x) = 0$).

(*b*) A series of line segments joining the points $(0,\frac{2}{3})$, $(1,\frac{2}{3})$, $(2,\frac{4}{3})$, a ray of slope -1 going left from $(0,\frac{2}{3})$, and a ray of slope 1 going right from $(2,\frac{4}{3})$. f is continuous everywhere and differentiable everywhere except at the vertices, that is, $x = 0, 1, 2$. $f(c) = \sum |x_i - c|\, p_i = \sum_{x_i>c} (x_i - c)p_i - \sum_{x_i<c} (x_i - c)p_i$. Thus $f'(c) = -\sum_{x_i>c} p_i + \sum_{x_i<c} p_i = P(X < c) - P(X > c)$.

(*c*) f is a minimum when $0 \leq x \leq 1$, where it has the value $\frac{2}{3}$.

(*d*) Proceed as in the solution to (*b*).

(*e*) Consider the intervals $I_0 = \{x < x_1\}$; $I_k = \{x : x_k < x < x_{k+1}\}$, $k = 1, \cdots, n-1$; $I_n = \{x > x_n\}$. g' is defined on each of these intervals and is greater on each successive one. Thus it is zero on at most one of these intervals. If $g'(c) = 0$, $c \in I_k$, then the minimum is attained precisely when $x_k \leq x \leq x_{k+1}$. Note that $\sum_{i=1}^{k} p_i = \frac{1}{2} = \sum_{i=k+1}^{n} p_i$. If $g'(c) \neq 0$ on all I_k, then there is a first k such that $g'(c) > 0$ on I_k. The minimum is then at x_k. We have $\sum_{i=1}^{k-1} p_i < \frac{1}{2} > \sum_{i=k+1}^{n} p_i$.

(*f*) Let $x_1 = 1$, $x_2 = 2$, $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$. Then the median (interval) is the single point 2. But $E(X) = \frac{5}{3} \neq 2$.

4.8 If X_i is the total for the ith die, then $X = X_1 + \cdots + X_n$ and the X_i are mutually independent. $E(X_i) = (1 + \cdots + 6)/6 = \frac{7}{2}$ and $E(X) = \Sigma E(X_i) = 7n/2$. $V(X_i) = E(X_i^2) - E(X_i)^2$. $E(X_i^2) = (1^2 + \cdots + 6^2)/6 = \frac{91}{6}$ and $V(X_i) = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$. Since the X_i are pairwise independent, $V(X) = \Sigma V(X_i) = 35n/12$.

4.9 $E([X - E(X)]^3) = E(X^3) - 3E(X^2)E(X) + 3E(X)E(X)^2 - E(X)^3 = E(X^3) - 3E(X^2)E(X) + 2E(X)^3$.

$$\begin{aligned} E(X^3) &= E([(X - E(X)) + E(X)]^3) \\ &= E([X - E(X)]^3) + 3E([X - E(X)]^2)E(X) + E(X)^3. \end{aligned}$$

4.13 (*a*) $4n/5 + 5$

(*b*) $2\sqrt{25-n}\,/5$.

(*c*) $-1/4$ per error.

4.14 (*a*) Unanswerable. Insufficient information is given in the problem.

5.1 Let S be the positive integers and let $P(n) = 2^{-n}$ for each n. Let $X(n) = 2^{n^2}$. Then $E(|X|^r) = \Sigma 2^{n^2r-n}$ and when nr exceeds 1, as it will for any positive r and all sufficiently large n, the terms each exceed 1. Thus the series diverges and the rth absolute moment is not defined.

5.2 The first condition on the range is necessary according to the text, page 79 lines 3*, 2*. The necessity of the second condition follows from calculating $E(X)$ using the first condition. A direct calculation verifies that the conditions are sufficient.

5.3 Let S be the positive integers, $P(n) = 2^{-n}$, $X(n) = n$. Then $E(|X|^r) = \Sigma 2^{-n} n^r$. It suffices to prove that this series converges. It is in fact true that $\Sigma x^{-n} n^r$ converges for all x with $|x| < 1$. Letting $t_n = x^{-n} n^r$ and applying the ratio test, we have $t_{n+1}/t_n = x(1 + 1/n)^{-r}$, which tends to x as n increases.

5.5 Let $f(p) = p(1 - p)$. Then $f'(p) = 1 - 2p$ and $f''(p) = -2$ so there is precisely one maximum, at $p = \frac{1}{2}$, where $f(p) = \frac{1}{4}$.

Alternately, we may "complete the square." The following expressions are equivalent: $p - p^2 \le \frac{1}{4}$, $4p - 4p^2 \le 1$, $-1 + 4p - 4p^2 \le 0$, and $-(1 - 2p)^2 \le 0$. Since the last inequality is true, it follows that the first is also.

5.6 (*a*) Let the polynomials of degree n be $P(x)$ and $Q(x)$ and let $P(x_i) = Q(x_i)$ for $x_1 < \cdots < x_{n+1}$. Then $P(x_i) - Q(x_i) = 0$ so $P(x) - Q(x)$ is a polynomial with $n + 1$ distinct roots. Since $P(x) - Q(x)$ is a polynomial of degree at most n, this is possible if and only if $P(x) - Q(x) \equiv 0$, that is, $P(x) \equiv Q(x)$.

(*b*) A simple set of points to choose are $-1, 0, 1, 2, 3$.

(*c*) $k^4 = (k)_4 + a_3(k)_3 + a_2(k)_2 + a_1(k)_1$. Cancelling out a common factor of k, expanding the terms on the right side, and equating coefficients of like powers of k gives the following equations for determining the a_i: $-6 + a_3 = 0$ hence $a_3 = 6$; $11 - 3a_3 + a_2 = 0$ hence $a_2 = 7$; $-6 + 2a_3 - a_2 + a_1 = 0$ hence $a_1 = 1$.

5.9 (*a*) $E(X^2) = \sum_{|X| \ge a} X(s)^2 P(s) + \sum_{|X| < a} X(s)^2 P(s) \le P(|X| \ge a) + a^2 P(|X| < a) \le P(|X| \ge a) + a^2$.

(*b*) Letting $X = S_n/n - \frac{1}{2}$, we have $E(X^2) = 1/4n$ and $P(|X| \ge 0.05) \ge 1/4n - \frac{1}{400}$. Since the right side is positive if and only if $n < 100$, this is the only time the inequality gives nontrivial information, with the above choice of X.

It is also true that $|X| \le \frac{1}{2}$ above so we can use $2X$ in place of X. Since $E((2X)^2) = 1/n$, we have $P(|X| \ge 0.05) = P(|2X| \ge 0.10) \ge 1/n - \frac{1}{100}$, four times the previous result.

5.11 n is sufficient if $[3(pq)^2 n^2 + (pq - 6p^2q^2)n]/(0.05n)^4 \le 0.01$. This simplifies to $n^3 - 3 \times 10^6 n - 2 \times 10^6 \ge 0$. Neglecting the constant term, we see that the root $n_0 = \sqrt{3}\, 10^3 \doteq 1{,}732.1$ of $n^3 - 3 \times 10^6 n = 0$ makes the left side negative, but that a slightly larger n should work. Substituting $n = n_0 + h$, we have after simplification the requirement $9 \times 10^6 h - 2 \times 10^6 \ge 0$ or $h \ge 2/9$. Therefore $n \ge 1733$ suffices. Thus,

the use of the fourth moment about the mean gives a considerably better estimate in this example than did the use of the variance.

7.1 The problem is to show that $\Sigma n^r c^n/n!$ converges for each $r \ge 0$. Letting $a_n = n^r c^n/n!$, we have (intending to use the ratio tests) $a_{n+1}/a_n = (1 + 1/n)^r c/(n + 1)$. For fixed r, this approaches $c/(n + 1)$ as n increases. But c is fixed so this tends to zero. Thus the ratio test shows the series converges.

7.2 (*a*) In talking of $P_n(E)$, it is understood that E is a nonempty subset of $0, 1, \cdots, n$. Consider the elementary event $E = \{n\}$. Then

$$(P(E) - P_n(E))/P_n(E) = f(n)/f_n(n) - 1.$$

The ratio $f(n)/f_n(n) = n^n/n!e^t \ge n/e^t$, since $n! \le n^{n-1}$. The latter increases without bound so the relative error does also.

(*b*) Suppose E is not an elementary event. Then we can write $E = E_1 + E_2 + \cdots + E_k$ with the E_i as elementary events. It is understood that $E \subset \{0, 1, \cdots, n\}$. If $P(E_i) = a_i$ and $P_n(E_i) = b_i$, the relative error for E is $\Sigma a_i/\Sigma b_i - 1$ and for the E_i it is $a_i/b_i - 1$. The problem reduces to establishing the two inequalities $\Sigma a_i/\Sigma b_i \le \max(a_i/b_i)$ and $\Sigma a_i/\Sigma b_i \ge \min(a_i/b_i)$. It further suffices to consider the case $k = 2$ for the general case can then be established by induction.

To establish $(a_1 + a_2)/(b_1 + b_2) \le \max(a_1/b_2, a_2/b_2)$, we may suppose that $a_1/b_1 \ge a_2/b_2$. Then $a_2 \le a_1 b_2/b_1$ and $(a_1 + a_2)/(b_1 + b_2) \le a_1/b_1 = \max(a_1/b_1, a_2/b_2)$. The other inequality follows from using $a_1 \ge a_2 b_1/b_2$ and substituting for a_1.

7.3

$$E(X) = \Sigma k f(k) = e^{-c}\Sigma c^k/(k - 1)! = c.$$
$$E(X^2) = \Sigma k^2 f(k) = \Sigma k(k - 1)f(k) + \Sigma k f(k) = c^2 + c.$$

Thus $V(X) = E(X^2) - c^2 = c$.

7.5 The Poisson approximation with $np = 1$ gives the probability of 7 coming up three or more times as

$$e^{-1} \sum_{k \ge 3} 1/k! = e^{-1}(e - 1 - 1/1! - 1/2!) = (e - 2.5)/e \doteq 8.05\%.$$

Chapter 4

3.1 In order to break even or better, we must have $S_n \ge n/36$. Since $np = n/37$ and $\sqrt{npq} = 6\sqrt{n}/37$, this is equivalent to $S_n^* \ge \sqrt{n}/216 = 8.97$. Using the estimate cited, we have $1 - F(8.97) \doteq 10^{-19}$, still inconceivably remote.

3.2 We assume that adults with I.Q.s of $75 = 1200/16$ or more will be bright enough to watch the 12 year old's program and that people who are not adults, and adults of I.Q. below 75, will have no interest in the program.

This means that adults from $-25/16 = -1.56$ standard deviations on up, or (from Table 2) 94.1% of the total adult audience, will be potential viewers of this program. If the program were designed for adults with an M.A. of 16 or more, only 50% of the adults would be potential viewers. Thus the potential audience would only be $50/94.1 = 53\%$ as large.

3.3 (*a*) We take $t = 6.25$ which yields $2.13 \times 10^{-10} < 1 - F(6.25) < 2.19 \times 10^{-10}$.

3.5 We want to find the greatest n such that $P(S_n > 1000) \leq 1\%$. when $p = q = \frac{1}{2}$. If we try $n = 2000$, we see at once that $P(S_n > 1000) \doteq 50\%$. Thus n is less than 2000. With $n = 2000$, $s \doteq 22$ so

$$P(S_n > 1000 + 2.4s) \doteq 1\%.$$

This suggests that we next try an n such that $n/2 + 2.4s \doteq 1000$, or $n = 2000 - 4.8s \doteq 1894$. If $n = 1894$ (this has a smaller s, as we can see at once, so it will be a little too small), $s = 21.76$ and $P(S_n > 1000) = 1 - F(2.46) = 0.69\%$, and $P(S_n > 997) \doteq 1\%$. With $n = 1900$, $s = 21.79$ and $P(S_n > 1000) = 1 - F(2.32) = 1.02\%$. Thus (by interpolation) $n = 1899$ is the solution, for $P(S_{1899} > 1000) \doteq 0.97\%$. Clearly this crude procedure can be formalized into an elegant and efficient process.

Index

A posteriori probability, 41
A priori probability, 41
Abbott, Edwin, 47
Absolute moment, 74
  first, 77
Ace-six flats, 47, 72
Additivity, countable, 52
  of probability measure, 30
Admission procedure, 123
Alpha particles, 94
Antisymmetry, of set inclusion, 21
Archimedes, 73
Argument, 56
Arithmetic mean, 67
Associative law, 22
Astragals, 5
Average, 67

Base, 102
Bayes' rule, 40 ff
  causes in, 41
  in discrete sample spaces, 53, 54
  in finite sample spaces, 41
  hypotheses in, 41
Bell, E. T., 1 fn, 11, 12, 26, 47
Bernoulli trials, 46, 80, 81, 82, 84, 89, 123
  fourth moment of, 82
  infinite sequence of, 85
  runs in, 72
Bernstein polynomials, vi
Bertrand's paradox, 100–102
Bets, 85
Binary operation, 22
Binomial coefficients, 10
  computation of table of, 12
Binomial distribution, 69, 89
  see also Normal approximation and Poisson approximation
Binomial random variable, 74
Binomial theorem, 1, 10
  proof of, 17
Birkhoff, Garrett, 18, 26
Birthday problem, 36
Blackjack, cheating in, 42
  dealer's probabilities in, 13
Boltzmann's constant, 123
Boole, George, 25 fn
Boolean canonical forms, 25
Boolean equations, 25
Borel, Emile, 39
Borel sets, 103 ff
Bortkiewicz, 94
Bounded set of real numbers, 71, 74, 78
Bridge, 10, 12, 13, 32, 39
  honor cards, 64
  void in, 49
  Yarborough in, 32
Brownian motion, 122
Buffon's needle experiment, 99–100, 103
Bush, Abelson, and Hyman, v

California short-form test, 122
Card shuffling, 48
Cardano, Gerolamo, 3, 3 fn
Cartesian product, 43 ff
  of infinitely many discrete spaces, 85 ff, 104
  as joint outcome of n experiments, 45

Cartesian product, probability measure
 for, 45–46
 of subsets, 44
Castor, 31
Cauchy distribution, 107
 principal value, 108, 109
Center of gravity, 73, 107
Central limit theorem, 121
Characteristic function, 29, 106, 111
Chebychev's inequality, 79 ff
Cheron, A., *see* Borel, Emile
Chevalier de Méré, 3
Coin, biased, 30
 guessing outcomes, 46
Coin tossing, 50, 118
 runs in, 54
Combinations, number of, 10
Combinatorial analysis, 7 ff
Commutative law, 22
Complement, 21
 relative to a subset, 23
Composition of functions, 63
Conditional convergence, 68
Conditional probability, 36 ff
 in discrete sample spaces, 53
Conditionally convergent integral, 109
 series, 68
Constant random variable, 73, 77
Continuity in mathematics, 97–99
Continuous density functions, 103 ff
 random variable, 105
Continuum, 103
Correlation coefficient, 76
Countably infinite set, 51
Counterimage, 57
Courant, R., 42
Covariance, 75, 78
Cover, 22, 23, 52
Covering, 22, 23
Craps, casino game, 8, 72
 don't pass, 72
 expectation, 72
 field, 54
 home game, 72
 pass, 72
 point, 72

de Morgan's laws, 25, 52
Density function, continuous, 103 ff
 discrete, *see* Frequency function
 piecewise continuous, 105
Dependent random variables, 62
Dependent variable, 57
Descartes, René, 43
Determinism, 2, 3
Dice, Ace-six flats, 47, 72
 crooked, 47, 73
Difference of sets, 23
Diffusion constant, 122
Discrete sample spaces, 50 ff
 extension to, of finite sample space
  results, 53
Disjoint sets, 22
 mutually, 52
 pairwise, 52
Distribution function, continuous, 105
 discrete defined differently, 65
 general, 124
 jump in, 124
Distributive law(s), for sets, 24, 27, 52
 left, 27
 right, 27
Divine calculator, 2
Dummy variable, 14, 65

e, 96
Einstein, Albert, 3 fn, 122
Element of a set, 21
Equally likely alternatives, 4 ff
Equiprobable measure, 30, 54
Error, experimental, 2, 6
Essentially constant, 57
 equal, 57
Event, certain, 20
 compound, 4
 elementary, 4
 impossible, 7, 20, 73
Expectation of random variable(s), bi-
  nomial, 69, 70
 continuous, 106
 discrete, 66 ff
 of a function of, 70
 interpreted, 67
 of sums, 69, 76
 undefined, 108
Experiment, 4

Exponential distribution, 109, 110
Exponential function, 96
Exponentiation, 27

Feller, William, 4, 11, 26, 35, 39, 49,
 83, 95, 114, 139
Fermat, Pierre, 3, 3 fn
Finite sample space, 20
First moment, 74
Fixed fraction betting, 84
Frank, C. J., vii
Frequency function, discrete, 60 ff
 axioms for, 60
 binomial, 60
 illustrated, 61
Frequency, relative, 33, 34
 and conditional probabilities, 36
Function(s), 27 ff
 biunique, 28
 characteristic, 29, 106, 111
 composition of, 63
 constant, 28
 domain of, 27
 domain space of, 27
 evaluation at *x*, 28
 even, 76
 extension of, 28
 factorial, 109, 110
 graph of, 59
 inverse of, *see* Inverse of a function
 many-to-one, 28
 many-valued, 28
 max of, 29
 of *n* variables, 43
 odd, 120
 one-to-one, 28
 onto, 28
 product of, 29
 range of, 27
 range space of, 27
 rational, 110
 single-valued, 28
 step, 111, 112
 synonyms for, 27

Galton, 75 fn
Gambling systems, proof that classical
  ones fail, 85
Gambling systems, winning, 85
Geiger, 94
Gelbaum, Bernard R., vii
Goldberg, 25, 26, 39, 42, 137
Gould, Sidney H., 3 fn
Graphs, 59

Heads or tails guessing, 46
l'Hospital's rule, 109

Independence, denial of, 81
 depends on measure, 40
 justified, 81
 mutual, for random variables, 62
 mutual, for sets, 38, 53
 pairwise, for random variables, 61
 pairwise, for sets, 38
Independent events, 37, 46, 53
 random variables, 61, 62, 77
 sets, 38
 trials, 45, 85
 variable, 57
Induction, mathematical, 16
 double, 18
 limitations of, 18
Integral test, 82
Integrals, absolute convergence of, 108
 Cauchy principal value of, 108
 conditionally convergent, 109
 negative part, 108
 positive part, 108
Intelligence quotient, 119–120
Intersection of sets, arbitrary, 52
 finite, 22
Interval, 98
 unit, 103
Intuition, probabilistic, 6
Inverse image, 57
Inverse of a function, 57 ff
 associated function of, 59
 domain of, 59
 illustrated, 58
IQ, *see* Intelligence quotient
Irrational numbers, 102

Jump, 124

Kemeny, 11, 26
Kolmogorov's inequalities, 83, 84
Kuller, Robert G., vi

Laplace, Pierre Simon, 2, 4, 113
divine calculator, 2
Law of averages, *see* Law of large numbers
Law of large numbers, 2, 67
see also Strong law of large numbers and Weak law of large numbers
Leibnitz, 1 fn
Lepley, William, 46
Lognormal distribution, 122

MacLane, Saunders, *see* Birkhoff, Garrett
Markov chains, vi
Mass, center of gravity of, 73
distribution, 73
moment of, 73
moment of inertia of, 74
Matching, 48
Mathematical induction, 16 ff
double, 18
limitations of, 18
Mathematical model, 6
independence in, 81
for stock market, 64
Maxwell-Boltzmann law, 123
Measure, probability, product, 88
uniform, 106
see also Probability measure
Measure theory, 123
Measure zero, set of, 57
Measurement, 55
Median, 78
Misprints, 96
de Moivre, 113
Moment generating functions, vi
Moment of inertia, 74, 107
Moments, continuous random variable, 106, 110
discrete random variable, 73 ff
discrete density function, 79
fourth, 82, 84
de Morgan's laws, 25, 52

More, Louis T., 1 fn
Natural logarithms, base of, 96
Newman, James R., 4
Newton, Sir Isaac, 1 fn
binomial theorem discovered by, 1 fn
law of universal gravitation, 1
mechanics, 1, 2
Newtonian mechanics, 1
Neyman, J., vi, 113
Nondiscrete sample space, 98
Normal approximation, 111 ff
computer study of, 123
error of, 113
illustrated, 111, 112
Normal density function, 113
Normal distribution, 89, 110 ff
illustrated, 116, 117
tables, 115, 116
Null set(s), 57
countable union of, 110

Odd function, 120
Ordinary moment, 74, 75
Origin, of coordinates, 43

Partition, 22, 23, 52
Parzen, Emmanuel, 35, 39, 42, 49, 95, 114
Pascal, Blaise, 3, 3 fn
Pascal's triangle, 12
Permutation, 48
Phenomena, analysis of, 2
chance part of, 2
deterministic part of, 2
Piecewise continuous density function, 105
Plane geometry, as model of mathematical system, 5, 6, 19, 80
Poisson approximation, 89 ff
computer study of, 95
distribution, 78, 89 ff
errors in, 89–91, 95
Poker, 12, 40
equivalence classes of hands, 13
types of hands, 13
Polls, 20
Pollux, 31

Polynomial, as density function, 110
Population, 9
individuals of, 9
Primitive terms, 19
Probability, a posteriori, 41
a priori, 41
of causes, 41
conditional, 36 ff
experimental determination of, 34
law, 106
Probability measure, 30 ff
additivity of infinite sample space, 30
on discrete sample spaces, 52
equiprobable, 30
on finite sample spaces, 30 ff
general, 104
non-negativity of, 30
normality of, 30
uniform, 106
Probability space, finite, 31
realization of, 35
refinement of, 32
Procedure for integration, 109
Product, finite, 15
Product measure, 88
Product set, 86
Proposition, in induction, 16
Prussian army, death by horsekick in, 92
Pseudorandom numbers, 35

Quantum theory, 2
Queues, theory of, 92

Radioactive decay, 2, 50, 93–94, 107
Random sample, 10, 39
Random variables, 55 ff
bets as, 85
constant, 56, 73, 77
continuous, 105
dependent, 61
expectation, 66 ff
finite range, 78
independent, 61, 77
moments of, 73 ff
mutually independent, 62
pairwise dependent, 61

Random variables, pairwise independent, 61
product of, 56
scalar multiple of, 56
standardized, 76, 77
sum of, 56
Rational numbers, 51, 97, 102
Raw scores, 119
Real numbers, bounded set of, 71, 74, 78
complete, 97
as a line, 97
unbounded set of, 71
uncountable, 86, 89
Rectangle, 87
Reflexivity of set inclusion, 21
Relative frequency, 33, 34
and conditional probabilities, 36
justification as interpretation of probabilities, 80, 81
Repeating decimals, 97, 102
Robbins, H., vi
Roulette, 39, 95, 118
bets offered, 32
biased wheel, 34
expectation in Nevada, 66, 67
fair wheel, 34
Run(s), expected length of, 72
first, 72
second, 72
Rutherford, 94

Sample, 8
random, 10, 39
Sample space, 20
discrete, 50 ff
finite, 20
Sampling, 8, 20, 36
ordered, 9, 36
random, 10
unordered, 9, 36
with replacement, 8, 9, 36
without replacement, 9, 36
Schleifer, *see* Kemeny
Second moment, 74
Series, absolute convergence, 67
conditional convergence, 68
convergence, 50

Series, exponential, 50, 96
  geometric, 50, 51
  non-negative, 50, 51
  rearrangement of, 50, 68
Set, bounded, 71
  complement of, 21
  element of, 21
  empty, 20
  finite, 20, 21
  infinite, 20, 51
  null, 57
  of subsets, 21
  synonyms for, 19
  unbounded, 71
  universal, 20
Set theoretic operations, 24
  interpreted in terms of events, 24
Set theory, elementary, 19 ff
Sigma field, 88, 89, 103 ff
Silver Slipper casino, closed for cheating, 47
Snell, *see* Kemeny
Standard deviation, 74
Standardized random variable, 76, 77
Stars, distribution of, 92–93
Statistical inference, 40
Step functions, 111, 112
Stirling's formula, vi
Stock market, 64, 121
Strong law of large numbers, 81, 82, 87
  extended, 85
Subpartition, 22, 52
Subpopulation, 40
Subset, of one element, 21
  proper, 20
Sum, infinite, telescoping, 15
  see Series
Summation, finite, 14 ff
  dummy variables in, 14, 65
  limits of, 14
Szabo, Alois, 118

Tails of a distribution, 79, 80
Thompson, *see* Kemeny
Thorp, Edward O., 12, 14, 46, 85
Todhunter, 3
Transitivity of set inclusion, 21
Trees, vi
Trials, 33

Unary operation, 22
Unbounded set of real numbers, 71
Uncountable sample space, 98
Uncountably infinite set, 51, 86, 89
Uniform measure, 106
  probability law, 106
Union of events, probability of, 47 ff
Union of sets, arbitrary, 52
  associativity of, 22
  commutativity of, 22
  finite, 21
Unit interval, 103

Vaccine testing, 92
Variable, 56
  dependent, 57
  dummy, 14, 65
  independent, 57
Variance, 74
Vector space, 57
Venn diagrams, 22–24
Voting coalitions, 26
  power, 26

Weak law of large numbers, 81
  extended, 84
Weierstrass' theorem, vi
Wilson, Allan, 34, 70, 85

Yarborough, Duke of, 32
  hand at bridge, 32

Zeroth moment, 74