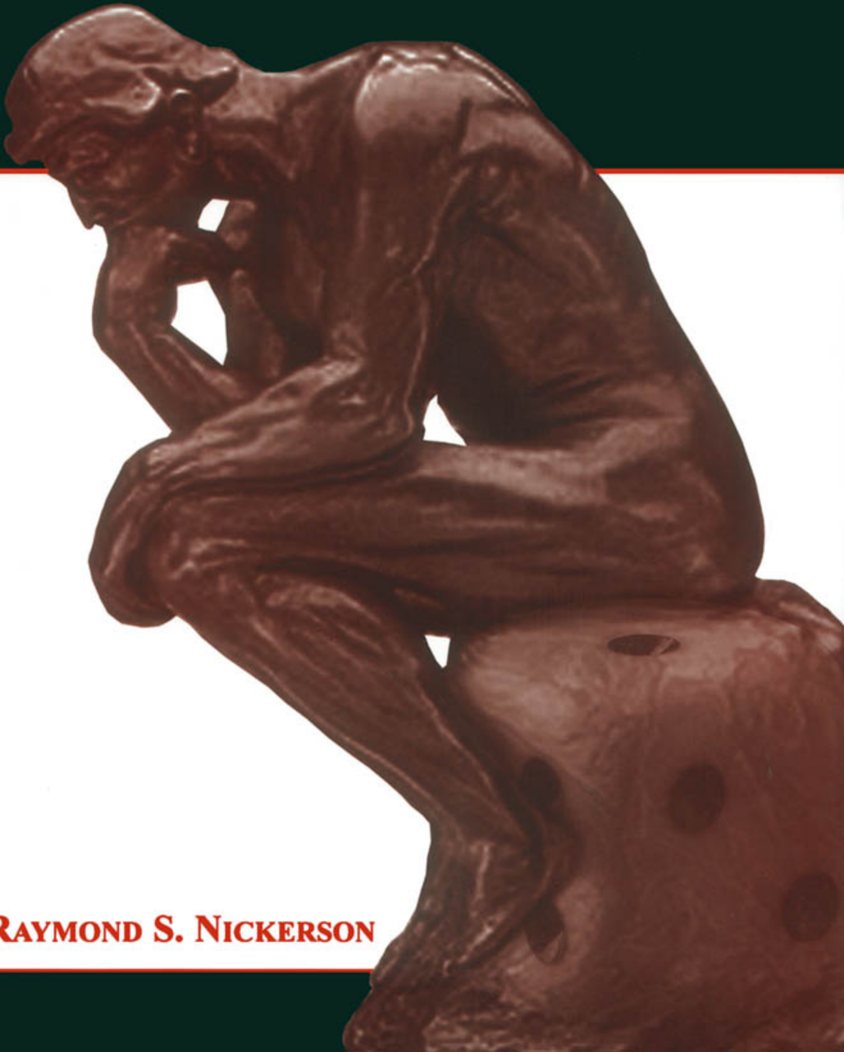
Cognition and Chance

The Psychology of Probabilistic Reasoning



Raymond S. Nickerson

Cognition and Chance

The Psychology of Probabilistic Reasoning

This page intentionally left blank

Cognition and Chance

The Psychology of Probabilistic Reasoning

Raymond S. Nickerson
Tufts University

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2004 Mahwah, New Jersey London

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

p. cm.
Includes bibliographical references and index.
ISBN 0-8058-4898-3 (cloth : alk. Paper)
ISBN 0-8058-4899-1 (pbk. : alk. paper)

2003
—dc22 2003000000
CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

To the fond memory of
Carlton Gregory
and
A. Douglas Glanville
exceptional teachers
who honored their students by making them think

This page intentionally left blank

Contents

Preface

Everybody thinks probabilistically, whether knowingly or not. We judge the likelihood that the barking dog will bite, that the rumor that the company is about to have a big layoff is true, that bottled water is purer than what one gets from one's kitchen faucet ... The ability to think probabilistically is important for many reasons. Lack of it makes one prone to a variety of irrational fears and vulnerable to scams designed to exploit probabilistic naiveté, precludes intelligent assessment of risks, ensures the operation of several common biases, impairs decision making under uncertainty, facilitates the misinterpretation of statistical information, precludes critical evaluation of likelihood claims, and generally undercuts rational thinking in numerous ways.

Often we lack the kind of evidence on complex issues that would permit us to draw a conclusion that we can be certain is correct. Frequently we have to make decisions on the basis of incomplete information and we cannot be sure of their consequences. But the need to settle for incomplete and uncertain information does not mean that our reasoning and decision making must be reduced to pure guesswork. Usually information of a statistical or probabilistic sort is available, or at least there is a basis for making some assumptions about the statistical or probabilistic characteristics of a situation of interest. One who can use this type of information effectively should do better, on the average, than one who cannot.

How good are individuals at thinking probabilistically? How consistent is people's reasoning under uncertainty with the principles of probability theory and mathematical statistics? These questions have been of considerable interest to researchers and the literature on this topic is very large. The evidence that has been produced is mixed. On the one hand are numerous indications that certain basic principles are poorly understood and that reasoning that should make use of those principles often is faulty. On the other hand are some experimental results that support the view that in many circumstances people's intuitions about statistics or probabilistic events are quite good.

My own interest in probabilistic reasoning derives in part from the belief that we all engage in it, more or less constantly, that we sometimes reason probabilistically in ways that suit our purposes very well and that we sometimes do rather poorly in this regard, and, finally, that a better understanding of probability and our abilities and limitations as probabilistic thinkers might help improve our thinking generally. But even if one could not legitimately claim that there are compelling practical reasons for studying probability and probabilistic thinking, I suspect that the allure of the topic, for me at least, would remain. I find probability, chance, randomness, and closely related concepts fascinating. I do not pretend to know what these terms mean in any deep sense; in fact the more I have thought about them, the more elusive whatever their referents are have become. I know that I am not alone in this respect and, indeed, believe that I am in rather good company.

My motivation for writing this book, as nearly as I can tell, was to learn something about probabilistic reasoning—about its origins, its status, its use in various contexts, what recent research has revealed about it, and so on. One should write a book, I have been told, with a specific audience in mind. This is undoubtedly good advice, but I confess to not following it. What I have written has been determined primarily by what I have found interesting. The book is divided roughly into two parts. The emphasis in the first part, which includes the first seven chapters, is historical and conceptual. The later chapters focus more on what research has shown about people's abilities and limitations as probabilistic thinkers.

Chapter 1 presents the history of the development of probabilistic ideas beginning with the seminal correspondence between Blaise Pascal and Pierre de Fermat in 1654 and continuing with the contributions of other 17th- and 18th-century figures. The concept of randomness, which is at once elusive and central to probability theory, is the subject of chapter 2. Coincidences and the question of what makes them interesting are discussed in chapter 3. The notion of inverse probability, as represented by Bayes's theorem, and Bayesian decision making and reasoning are discussed in chapter 4. Chapter 5 presents a variety of problems that illustrate some of the subtleties that can be involved in dealing with probabilities. Chapter 6 continues this theme with a discussion of several paradoxes and dilemmas involving probabilities. Chapter 7 gives an account of the birth of statistics and its increasing applications in science.

Chapter 8 reviews empirical work on the question of how well people are able to estimate probabilistic variables and predict the outcomes of probabilistic events. Chapter 9 continues the review of empirical work, now with a focus on people's abilities to perceive covariation and contingency, and the factors that affect their performance in this regard. Chapter 10 deals with decision making under uncertainty and the notions of expected utility, inductive heuristics, and probabilistic mental models. Chapter 11 focuses on the general question of how good people are as intuitive statisticians. Some concluding remarks comprise chapter 12.

I can only hope that some readers will learn as much from reading the book as I think I did from writing it, and with as much pleasure. As to who such readers might be, I hope they might include psychologists, philosophers, and economists for whom probabilistic reasoning is an essential aspect of their efforts to understand the workings of the human mind more generally, researchers studying probabilistic reasoning, and students preparing to do so. I believe the book could be used as a major text for a graduate or upper-class seminar on probabilistic reasoning, or as a supplementary text for graduate or upper-class courses in cognitive psychology, philosophy, economics, or any subject for which probabilistic reasoning is a key topic. I have tried to write also, however, for the general reader who has an interest in the subject and little or no technical training beyond standard high school math. Familiarity with basic algebra should suffice for the great majority, if not all, uses of mathematics, of which there are relatively few.

I owe thanks to Jonathan Baron, Susan Butler, Richard Chechile, Julie Downs, Ruma and Rafael Falk, R. Duncan Luce, Nathan Nickerson, Richard Shiffrin, Salvatore Soraci, Howard Wainer, and Michael Wenger for helpful comments on parts of the manuscript and/or stimulating conversations on one or another aspect of probabilistic reasoning, to my granddaughter Amara Nickerson for helping to organize and check references, and, as always, to my wife Doris for her constant love, companionship, and encouragement.

This page intentionally left blank

CHAPTER

1

Probability and Chance

Probability is ... the philosophical success story of the first half of the twentieth century.

—Hacking (1990, p. 4)

Strictly speaking it may even be said that nearly all our knowledge is problematical; and in the small number of things which we are able to know with certainty, even in the mathematical sciences themselves, the principal means for ascertaining truth—induction and analogy—are based on probabilities.

—Laplace (1814/1951, p. 1)

Subjective probabilities are required for reasoning ... a theory of partially ordered subjective probabilities is a necessary ingredient of rationality.

—Good (1983, p. 95)

BEGINNINGS

The use of chance devices and the drawing of lots for purposes of sortilege and divination were common to many cultures of antiquity. Classical Greek literature contains numerous references to games of chance at least as early as the Trojan wars, and there is evidence to suggest that such games were known in Egypt and elsewhere long before then. One of the earliest known written documents about the use of chance devices in gaming is in the Vedic poems of the

Rgveda Samhita. "Written in Sanskrit circa 1000 B.C., this poem or song, called the 'Lament of the Gambler,' is a monologue by a gambler whose gambling obsession has destroyed his happy household and driven away his devoted wife" (Bennett, 1998, p. 34).

Gambling had become a sufficiently common form of recreation in Europe by the time of the Roman Empire that laws forbidding it were passed—and largely ignored. The emperor Caesar Augustus (63 B.C.–14 A.D.) was an avid roller of the bones. The bones, in this context, were probably astragali, a form of dice made from the heel bones of running animals. The astragalus, sometimes referred to as a talus, huckle-bone, or knuckle-bone, was shaped in such a way that, when tossed, it could come to rest with any of four sides facing up, the other two sides being somewhat rounded. It is known to have been used in the playing of board games at least as early as 3500 B.C. Dice, which presumably are evolutionary descendants of astragali, are known to have existed in the Middle East as early as the third millennium B.C. According to Bennett (1998), the earliest known six-sided die was made of baked clay around 2750 B.C. and was found in what was once Mesopotamia and is now Northern Iraq.

In view of the fact that chance devices have been used for a variety of purposes for so long, scholars have found it difficult to explain why quantitative theories of randomness and probability were not developed until relatively recent times. Although why a theory of probability was so long in coming remains unknown, there has been much speculation as to what some of the contributing factors may have been. Hypothesized deterrents include pervasive belief in determinism, lack of opportunity to observe equiprobable sets (astragali did not turn up all four faces with equal relative frequency), absence of economic incentives, and lack of a notational system suitable for representing the critical ideas and computations. Each of these explanations has been challenged, if not discredited to one or another degree (Hacking, 1975). There appear what Gigerenzer et al. (1989) refer to as suggestive fragments of probabilistic thinking in classical and medieval literature, and there may have been some theoretical ideas about probability, especially in India, that are lost to us, but apparently nothing approaching a systematic theoretical treatment of the subject was attempted in Europe until the 17th century.

Gigerenzer et al. (1989) attribute considerable significance to the Reformation and Counter-Reformation and the associated clashes between extremist views on faith and reason, which influenced attitudes regarding how beliefs should be justified:

> Confronted with a choice between fideist dogmatism on the one hand and the most corrosive skepticism on the other, an increasing number of seventeenth-century writers attempted to carve out an intermediate position that abandoned

> all hope of certainty except in mathematics and perhaps metaphysics, and yet still insisted that men could attain probable knowledge.... The new criterion for rational belief was no longer a watertight demonstration, but rather that degree of conviction sufficient to impel a prudent man of affairs to action. For reasonable men that conviction in turn rested upon a combined reckoning of hazard and prospect of gain, i.e. upon expectation. (p. 5)

They continue:

> Mathematicians seeking to quantify the legal sense of expectations inevitably became involved in quantifying the new rationality as well. So began an alliance between mathematical probability theory and standards of rationality that stamped the classical interpretation as a 'reasonable calculus'; as a mathematical codification of the intuitive principles underlying the belief and practice of reasonable men. (p. 6)

Interest in probability was spurred by questions relating to games of chance, such as why the tossing of three dice turned a total of 10 more frequently than a total of 9. Much of the early thinking about the topic was prompted by the specific question of how to divide fairly the stakes in a prematurely terminated game of chance. The following problem appeared in Fra Luca Paccioli's *Summa de Arithmetic, Geometria et Proportionalità (The Whole of Arithmetic, Geometry and Proportionality),* published in 1494: "*A* and *B* are playing a fair game of *balla*. They agree to continue until one has won six rounds. The game actually stops when *A* has won five and *B* three. How should the stakes be divided?" (David, 1962, p. 37).

One solution by Paccioli was published in *De Divina Proportione (On Divine Proportion)* in 1509, and another by Gio Francesco Peverone in *Due Brevi e Facile Trattati: Il Primo d'Arithmetica, l'Altro di Geometria (Two Short and Easy Treatises: The First on Arithmetic, the Other on Geometry)* in 1558 (David, 1962). These early efforts considered the question of what would be a fair division in specific cases and did not attempt a general solution to the problem. The answers proposed for the cases considered differ from the answers that would be given to the same questions on the basis of probability theory as it now exists, employing, as they did, the notion of dividing the stakes in the same ratio as that of the games won, or a close derivative of it. It is doubtful whether either Paccioli or Peverone had a clear concept of probability. This type of problem appears repeatedly in the writings of 16th- and 17th-century mathematicians and differences of opinion as to the correct answer prompted heated debate (P. L. Bernstein, 1996; Todhunter, 1865/2001).

A significant early contributor to a theory of probability, as we know it—or to a theory of chance, as it was and is sometimes called—was Girolamo

Cardano, who wrote a book with the title *Liber de Ludo Aleae (The Book of Games of Chance)*. Although written, according to his own account, in 1525, it was not published until 1663. David (1962) credits Cardano with being the first mathematician to calculate a theoretical probability correctly. Galileo also wrote briefly about the numbers of ways different sums can be obtained with the tossing of three dice.

The Pascal–Fermat Correspondence

Blaise Pascal and Pierre de Fermat thought collaboratively about the stakes-division problem as a consequence of the question being put to Pascal by a gambler, Chevaleau de Méré. Pascal and Fermat exchanged letters on the topic over several months during the summer and fall of 1654. (The collaborators never met before, during, or after their remarkable correspondence.) The correspondence between Pascal and Fermat regarding the stake-division problem is worth considering in some detail, because it is viewed as one of the defining events in the emergence of probability theory as a mathematical discipline. And it provides a glimpse at the kind of struggle that the founders of probability theory had in trying to understand their own intuitions and to make them explicit and clear. An English translation of what is preserved of the correspondence—apparently the surviving record is not entirely complete— may be found in David (1962, Appendix 4).

The specific situation that Pascal and Fermat considered was this. Two players are involved in a series of games and play is terminated when one of the players, say A, is two games short of winning the series and the other player, say B, is three games short of doing so. How should the stakes be divided between the players at this point? The tacit assumption is that A should get a larger share than B because A would have been more likely to win the series had play continued, but exactly what proportion of the total should each player receive?

Fermat had proposed a way of analyzing the situation. In a letter dated Monday, August 24, 1654, Pascal repeats Fermat's analysis and argues that it works for two players but will not do so if more than two are involved. Fermat's analysis starts with a determination of the maximum number of additional games that would have to be played to decide the winner of the series. In this case, the number is four, because in the playing of four games either A will win two or B will win three. Fermat's method then calls for identifying all possible outcomes of four games, of which there are 16, determining the percentage of them that would be won by each player, and dividing the stakes in accordance with the ratio of these percentages.

The situation is represented in Pascal's letter by the following tabular arrangement in which each *column* represents one possible outcome of the four games, with *a* representing a win of an individual game by A and *b* a win by B

(Pascal actually used 1s and 2s in the last row of the table where I have As and Bs, but this is irrelevant to the point of the discussion):

a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b
a	a	a	a	b	b	b	b	a	a	a	a	b	b	b	b
a	a	b	b	a	a	b	b	a	a	b	b	a	a	b	b
a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
A	A	A	A	A	A	A	B	A	A	A	B	A	B	B	B

As this analysis shows, of the 16 possible outcomes of four games, A would win the series (by winning at least two individual games) in 11 cases and B (by winning at least three individual games) in 5. So, according to Fermat's reasoning, the stakes should be divided between A and B in the ratio 11 to 5.

Pascal argues that Fermat's analysis gives the correct answer, so long as there are only two players, but will not do so if there are more than two: "I must tell you that the solution of the problem of points for two players based on combinations is very accurate and true, but if there are more than two players it will not always be correct" (David, 1962, p. 240).

Before giving his reasons for believing Fermat's solution not to be reliable when there are more than two players, Pascal digresses to deal with an objection that a colleague, M. de Roberval, had raised to Fermat's solution, which Pascal had shown to him, even in the two-player case: "What is mistaken [according to de Roberval] is that the problem is worked out on the assumption that *four* games are played; in view of the fact that when one man wins *two* games or the other *three,* there is no need to play *four* games, it could happen that they would play *two* or *three,* or in truth, perhaps *four*" (David, 1962, p. 241).

In reporting to Fermat how he dealt with this objection, Pascal notes that he himself did not rely on the combinatorial method, "which in truth is not appropriate here," but that nevertheless he was able to construct an argument that the method gave the correct answer in this case. First he made the point that if the two players, finding themselves in the situation that one needed two games to win the series and the other three, agreed to play four additional games, then Fermat's analysis shows the correct division of stakes. De Roberval agreed with this, but denied that it would apply if the players were not compelled to play the four games.

Pascal then argued that the continuation of play after one or the other has won the series—after A has won two games or B three—can have no effect on the outcome, so whether or not the players do continue is irrelevant:

> Certainly it is easy to see that it is absolutely equal and immaterial to them both whether they let the game take its natural course, which is to cease play when

> one man has won, or to play the whole four games: therefore since these two procedures are equal and immaterial, the result must be the same in them both. Now, the solution is correct when they are obliged to play four games, as I have proved: therefore it is equally correct in the other case. (David, 1962, p. 242)

He does not say whether de Roberval was convinced.

Returning to the question of the applicability of Fermat's combinatorial analysis to cases in which there are more than two players, Pascal considers the following three-person situation. One player, whom I call A, needs one game to win and each of the other two, B and C, needs two. The maximum number of games that will be required to determine the winner is three, inasmuch as each of the 27 possible combinations of three wins contains either one win for A or two wins for either B or C. The problem, and the reason that Pascal dismisses the method, is that these are not mutually exclusive possibilities; some combinations contain more than one of them.

As before, Pascal represents the situation with a tabular arrangement in which each column identifies one possible outcome of three games:

a	a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b	b	c	c	c	c	c	c	c	c	c
a	a	a	b	b	b	c	c	c	a	a	a	b	b	b	c	c	c	a	a	a	b	b	b	c	c	c
a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
A	A	A	A	A	A	A	A	A	A	A	A	A			A			A	A	A	A			A		
				B						B		B	B	B		B						B				
								C									C			C			C	C	C	C

So, given the need of A to win one game, and that of the other two players each to win two, the playing of three games will not invariably produce an unambiguous result. Pascal dismissed the possibility of simply adding up the total "wins" for each player and dividing the stakes in the proportion 19:7:7 on the grounds that a combination that is favorable to two players (e.g., *a b b*) should not count as much for each of them as does a combination that is favorable to that player alone. One possibility that he considered is that of counting those combinations that are favorable to two players as half a win for each of them. This would give a division in the proportion 16:5.5:5.5.

Pascal argues that the latter solution would be fair if the players' intention was to play three additional games and to share the winnings equally if there happened to be two winners, but not if their intention was to play only until the first player to reach his goal had done so. He contends that, given the second intention, the solution is unfair because it is based on the false assumption that three games will always be played. A similar assumption caused no difficulty in the two-person situation considered earlier, but it is problematic here. The difference between the situations is that in the first one a combination with more than one winner could not occur whereas in this one it could. Pascal goes on to say that if play is to be continued only until one of the players reaches the number of games he needs, the proper division is 17:5:5. This result, he says, can be found by his "general method," which is not described in the letter.

In a reply to Pascal, dated Friday, September 25, 1654, Fermat agrees that the appropriate division in the three-person situation considered by Pascal is 17:5:5:

> I find only that there are 17 combinations for the first man and 5 for each of the other two: for, when you say that the combination *a c c* is favourable to the first man and to the third, it appears that you forgot that everything happening after one of the players has won is worth nothing. Now since this combination makes the first man win the first game, of what use is it that the third man wins the next two games, because even if he won thirty games it would be superfluous? (David, 1962, p. 248)

Fermat seems to be saying that if one uses his combinatorial analysis and credits a combination *only to the first of the players to reach his goal,* then one will get the right proportion in a straightforward way. If we do this with the previous table, for example, and let the top-to-bottom order of the rows represent the order in which the games are played, we get:

a	a	a	a	a	a	a	a	a	b	b	b	b	b	b	b	b	b	c	c	c	c	c	c	c	c	c
a	a	a	b	b	b	c	c	c	a	a	a	b	b	b	c	c	c	a	a	a	b	b	b	c	c	c
a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
A	A	A	A	A	A	A	A	A	A	A	A				A			A	A	A	A					
												B	B	B		B						B				
																	C						C	C	C	C

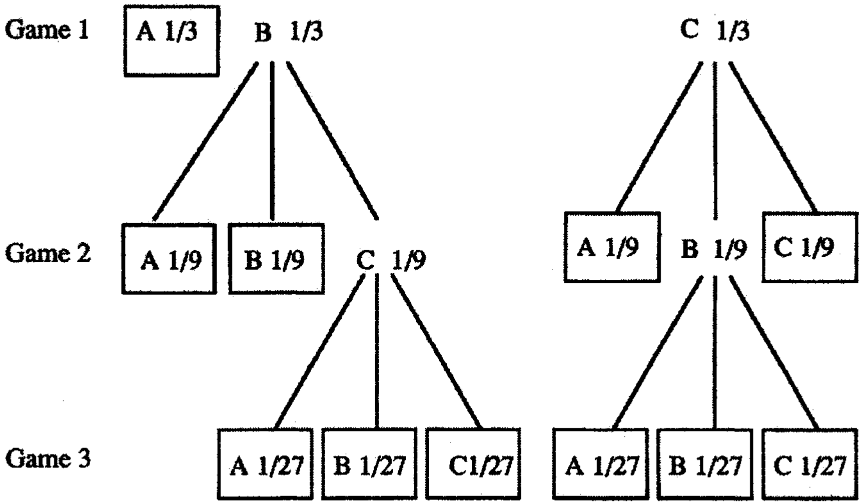
Fermat argues that this method is general in the sense that it works just as well no matter how many games are to be played. If it is done with four games, for example, it will be seen that A wins 51 of the 81 possible combinations and B and C each win 15, giving again the proportion 17:5:5. The tone of Fermat's letter suggests that he understood all this before Pascal pointed out the "problem" with his analysis of the two-person game.

Fermat proposes another way to view the situation. I quote his presentation of this view in its entirety, because it would be easy to change it subtly but materially by paraphrasing it:

> The first man can win, either in a single game, or in two or in three.
>
> If he wins in a single game, he must, with one die of three faces, win the first throw. [Throughout the correspondence Fermat and Pascal use the abstraction

> of an imaginary die with two or three faces.] A single [three-faced] die has three possibilities, this player has a chance of 1/3 of winning, when only one game is played.
>
> If two games are played, he can win in two ways, either when the second player wins the first game and he wins the second or when the third player wins the first game and he wins the second. Now, two dice have 9 possibilities: thus the first man has a chance of 2/9 of winning when they play two games.
>
> If three games are played, he can only win in two ways, either when the second man wins the first game, the third the second and he the third, or when the third man wins the first game, the second wins the second and he wins the third; for, if the second or third player were to win the first two games, he would have won the match and not the first player. Now, three dice have 27 possibilities: thus the first player has a chance of 2/27 of winning when they play three games.
>
> The sum of the chances that the first player will win is therefore 1/3, 2/9 and 2/27 which makes 17/27.
>
> And this rule is sound and applicable to all cases, so that without recourse to any artifice, the actual combinations in each number of games give the solution and show what I said in the first place, that the extension to a particular number of games is nothing but a reduction of the several fractions to a common denominator. There in a few words is the whole mystery, which puts us on good terms again since we both only seek accuracy and truth. (David, 1962, p. 248)

Although Fermat does not use it, his comments suggest the following tree representation of the situation he described—the case of three players, one of whom, say A, needs one more win and two of whom, B and C, both need two:



The boxed outcomes in this representation indicate possible terminal points. Play ceases after Game 1, for example, if that game is won by A, which will happen with probability 1/3. If B wins Game 1, which also will happen with probability 1/3, Game 2 is played. In 1/3 of *those* games A will win Game 2; so the probability that B wins Game 1 *and* A wins Game 2 is (1/3) × (1/3) or 1/9. And so on. The probability that A will win the series is the sum of the probabilities of the individual ways in which A can win, which is 1/3 + 2/9 + 2/27 = 17/27. The probability that B will win the series is 1/9 + 2/27 = 5/27, as is the probability that C will win it.

I have considered the correspondence between Pascal and Fermat at some length, because it provides a glimpse into the thinking of two giants of mathematics—Todhunter (1865/2001) calls them the most distinguished mathematicians of Europe at the time—as they attempted to deal with probabilistic concepts before the theory of probability had been developed to any great degree. (Fermat was not a mathematician by profession, but apparently did mathematics for the sheer pleasure that he derived from it. Extraordinarily productive, E. T. Bell [1937] calls him "the prince of amateurs" [p. 56].) One sees a progression of thought that involves a gradual increase in awareness of the various dimensions of the problem and the realization that convergence on a consensus requires making assumptions clear.

It is interesting to note that neither Pascal nor Fermat seemed intent on rushing into print to establish the priority of their early work in probability theory. As Todhunter (1865/2001) points out, it apparently was sufficient for each of these men to gain the approbation of the other. Todhunter notes, too, that the theory was advanced only little during the half century following their correspondence—Pascal soon turned his attention from mathematics, Fermat died in 1665, and there soon were other topics, such as the differential calculus of Newton and Leibnitz, to capture mathematicians' attention. Pierre de Montmort, writing shortly after the turn of the 18th century could claim, with only moderate exaggeration, to be exploring a subject that after being only slightly noticed had been entirely forgotten for sixty years (Todhunter, 1965/2001).

Montmort gives the following general solution to the "problem of points" that was the subject of the Pascal–Fermat correspondence. Let m and n represent the number of points that A and B, respectively, need to win the game, and let p represent the probability of A winning a single trial and q the probability of B doing so, $p + q = 1$. Let $r = m + n - 1$, the number of trials in which the game must be decided. A's chance of winning is given by

$$p^r + rp^{r-1}q + \frac{r(r-1)}{1\times 2}p^{r-2}q^2 + \ldots\ldots + \frac{r!}{m!(n-1)!}p^m q^{n-1},$$

and B's by

$$q^r + rq^{r-1}p + \frac{r(r-1)}{1\times 2}q^{r-2}p^2 + \ldots\ldots + \frac{r!}{n!(m-1)!}q^n p^{m-1}$$

This formula applies if $p + q = 1$, which is to say if one or the other player must win on each trial; a different formula is required if ties are allowed (Todhunter, 1865/2001, p. 97).

It is very easy, with the benefit of hindsight, to underestimate the difficulty with which some of the insights that we now take for granted were attained. The distinction between combinations and permutations, which is essential to an understanding of probability theory and is now encountered very early in any study of the subject, was obscured for a very long time. Bennett (1998) notes that for centuries learned people believed that the tossing of three dice had 56, not 216, possible outcomes, because they failed to make this distinction. She points out that although games involving the tossing of three dice had been popular since the days of the Roman Empire, the first known correct enumeration of all the equiprobable outcomes of the toss of three dice is attributed to Richard de Fournival, and was written sometime between 1220 and 1250.

Other Major Contributors to Theory Development

Another significant contributor to the early development of probability theory was Christiaan Huygens, who documented the collaboration between Pascal and Fermat in his *De Ratiociniis in Ludo Aleae (On Reasoning and Games of Chance)* (I have seen both 1654 and 1657 as the publication date), which, David (1962) says, served as the unique introduction to probability theory for the following half century. In this book, Huygens treated the question, as had Pascal and Fermat, of how to divide the stakes in an uncompleted game of chance, but he did so in a more general way. Here, also, he introduced the idea of mathematical expectation.

Another 17th-century work that contained some references to probability and a discussion of some ideas that might be considered precursory to statistical inferencing was *La Logique, ou l'Art de Penser (Logic, or the Art of Thinking)*, which was published in 1662, with financial support from Pascal, by a group at the Port-Royal monastery. Although the author of the book was not identified, Antoine Arnauld is believed to have been the primary, but not the only, one (P. L. Bernstein, 1996).

Noteworthy advances in the development of probability theory came early in the 18th century with such works as *Essai d'Analyse sur les Jeux de Hasard (Essay on the Analysis of Games of Chance)*, by Montmort (1708/1713), *Ars*

Conjectandi (The Art of Conjecture), by Jacob (sometimes Jakob, Jacques, or James) Bernoulli (1713), and *The Doctrine of Chances: or a Method of Calculating the Probabilities of Events in Play*, by Abraham de Moivre (1718; second and third editions of which were published in 1738, and, posthumously, in 1756), the last being an expanded version of *De Mensura Sortis (On the Measurement of Risks [or lots])* (1711).

The relationship between Montmort and de Moivre, a somewhat testy one, provides a glimpse at how egos can intrude even in the presumably dispassionate world of mathematics. De Moivre, born in 1667, was 11 years older than Montmort, but de Moivre outlived Montmort, who died in 1719, by 35 years. Because Montmort published his essay about 3 years before de Moivre published *De Mensura Sortis*, and de Moivre lived and produced much longer than Montmort, de Moivre is often considered Montmort's successor. Both contributed significantly to the advancement of probability theory, but neither was particularly generous in his assessment of the contributions of the other. Montmort was angered by what he took to be de Moivre's belittling of his work on probability theory, as put forth in *Essai d'Analyse sur les Jeux de Hasard*, when de Moivre referred to it in a condescending way in the preface to one of his own books on the subject. Montmort chastised de Moivre in print and de Moivre later acknowledged Montmort's contribution to the field. Montmort held that his essay contained implicitly all that was in de Moivre's *De Mensura Sortis* (Todhunter, 1865/2001).

Todhunter (1865/2001) contrasts the persons and work of Montmort and de Moivre this way:

> Montmort's work on the whole must be considered highly creditable to his acuteness, perseverance, and energy. The courage is to be commended which led him to labour in a field hitherto so little cultivated, and his example served to stimulate his more distinguished successor. De Moivre was certainly far superior in mathematical power to Montmort, and enjoyed the great advantage of a long life, extending to more than twice the duration of that of his predecessor; on the other hand, the fortunate circumstances of Montmort's position gave him that abundant leisure, which De Moivre in exile and poverty must have found it impossible to secure. (p. 134)

De Moivre made many contributions of unquestioned originality to the development of probability theory. His *De Mensura Sortis* was essentially a collection of 26 probability problems and solutions. In this regard, it is representative of much of the early work; most of the attention was directed at finding solutions to very specific problems. *The Doctrine of Chances* contains a more extensive collection of problems and solutions. Unlike *De Mensura Sortis,* however, *The Doctrine of Chances* (third edition) begins with a tutorial presentation of the subject with illustrations of the principles involved. Even so, many of problems subsequently discussed are likely to be abstruse to the reader who is not relatively adept at mathematics. David (1962) refers to the third edition of his *The Doctrine of Chances,* which was published posthumously in 1756, as "the first modern book on probability theory" (p. 171). Writing in the middle of the 19th century, Todhunter (1865/2001) credits de Moivre with contributing more to the theory of probability than any other mathematician up to that time with the sole exception of Laplace. Todhunter gives an extensive summary of the problems de Moivre considered.

Other notable mathematicians who were important to the early development of probably theory included other members of the Bernoulli family, Leonhard Euler, and Joseph Lagrange. Daniel Bernoulli published an essay titled "Specimen Theoriae Novae de Mensura Sortis" ("Exposition of a New Theory on the Measurement of Risk") in *Papers of the Imperial Academy of Sciences in Petersburg* in 1738, in which he distinguished among the concepts of value, price, and utility. P. L. Bernstein (1996) considers this paper to be "one of the most profound documents ever written, not just on the subject of risk but on human behavior as well" (p. 100).

Hacking (1975) credits Jacob Bernoulli with presenting, in his *Ars Conjectandi,* "the most decisive conceptual innovations in the early history of probability" (p. 143). The book, which was published eight years after Bernoulli's death—Bernoulli died before the fourth part of his four-part book was finished—contains the first limit theorem of probability, a key development in the history of probability theory. This theorem formalizes the idea, now known as the "law of large numbers," which Bernoulli considered to be intuitively compelling, that the relative frequencies of chance events are very close to the probabilities of those events in sufficiently large samples. The following statement of the theorem is from Uspensky (1937): "With the probability approaching 1 or certainty as near as we please, we may expect that the relative frequency of an event E in a series of independent trials with constant probability p will differ from that probability by less than any given number e > 0, provided the number of trials is taken sufficiently large" (p. 100).

Imagine an urn containing balls some proportion, p_w, of which are white. According to the limit theorem, if balls are drawn from such an urn randomly, one at a time, and replaced after each drawing, then as the number of drawings, *N,* increases indefinitely the probability that the proportion of white balls in the sample, *W/N,* gets arbitrarily close to the proportion of white balls in the urn. That is,

$$\lim (N \text{->}\infty)\ P(|\,p_w - W/N| < \varepsilon = 1, \text{ for any } \varepsilon.$$

Laplace (1814/1951) expressed the theorem this way: "The probability that the ratio of the number of white balls drawn to the total number of balls drawn does not deviate beyond a given interval from the ratio of the number of white balls to the total number of balls contained in the urn, approaches indefinitely to certainty by the indefinite multiplication of events, however small this interval" (p. 610).

Gigerenzer et al. (1989) refer to this result as a "curious mixture of the banal and the revolutionary" (p. 29). Banal because it expresses a relationship between number of observations and the confidence one should have in inferences drawn from them that, as Bernoulli had acknowledged, seems intuitively obvious, and revolutionary "because it linked the probabilities of degrees of certainty to the probabilities of frequencies, and because it created a model of causation that was essentially devoid of causes" (p. 29).

Bernoulli's limit theorem provides the basis for inferences in two directions. Assuming one knows the probability of the outcome of an event, one can invoke it to predict the approximate relative frequency of that outcome from a large number of such events. Conversely, if one does not know the probability of a particular outcome of an event, one can take the relative frequency with which that outcome occurs in a large number of such events as an approximation of the probability of that outcome. (The obvious circularity in these assertions is the basis for some of the puzzlement in discussions of what probability "really is.")

We have seen that a great deal of the early work on probability theory involved analyses of games of chance, generally played with cards or dice. Games of the day considered by various developers of the theory included Pharaon, Treize, Bassette, Her, Raffling, Hazard, Whist, and numerous others. Sometimes the problems considered involved imaginary dice—dice with some number of sides other than six. Progress involved correcting errors of earlier analyses, generalizing (extending an analysis of a game with a specified number of cards or specified number of rolls of a die to the general case; extending consideration of games with evenly matched players to consideration of games with players with different probabilities of winning), finding better representations of problems, proving conjectures, and replacing existing formulas or proofs with simpler or more elegant ones. Although important books were produced, much of the documentation of the thinking that was done and the progress that was made resides in correspondence among many of the major contributors—the Bernoullis, Montmort, Leibnitz, and so on. Numerous excerpts from this correspondence may be found in Todhunter (1865/2001).

Despite the fact that several eminent figures in the history of mathematics gave some attention to probability theory during the latter part of the 17th century and the early part of the 18th, it did not become a major focus of mathematicians more generally until much later. Perhaps the next noteworthy work published on the topic that is still referenced today was Pierre Laplace's *Théorie Analytique des Probabilités (Analytic Theory of Probabilities),* the first edition of which appeared in 1812, more than 150 years after the work of Pascal and Fermat, and remained the dominant publication on the subject for about a century (Kamlah, 1987). Todhunter (1865/2001) says of Laplace that the theory of probability was more indebted to him than to any other mathematician. The first edition of *Théorie Analytique des Probabilités,* but not the second and third editions, which were published in 1814 and 1820, contained a dedication to Napoléon-le-Grand. Laplace also published, in 1814, *Essai Philosophique sur les Probabilités (A Philosophical Essay on Probabilities)* for more popular consumption.

Laplace (1814/1951) referred to the theory as "at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it" (p. 196). Laplace recognized the importance of probability theory—"It is remarkable that a science, which commenced with the consideration of games of chance, should be elevated to the rank of the most important subjects of human knowledge" (p. 195)—and contended that if we considered its remarkable aspects, which he mentions, "then we shall see that there is no science more worthy of our meditations, and that no more useful one could be incorporated in the system of public instruction" (p. 196). However, after his work, interest in probability among mathematicians again waned and remained at a low level for the remainder of the 19th century and the first couple of decades of the 20th, despite the fact that its application was proving to be very useful, even in theoretical physics.

Kac (1964) attributes the relative lack of interest in probability theory among mathematicians immediately following Laplace to a feeling within the discipline that the theory was "built on loose and nonrigorous foundations" (p. 96). Laplace's definition of probability, he contends, is circular because it invokes the notion of equally likely outcomes, which notion is itself a probabilistic one. Kac notes too that "the field was plagued with apparent paradoxes and other difficulties. The rising standards of rigor in all branches of mathematics made probability seem an unprofitable subject to cultivate" (p. 96). There were a few notable 18th- and early-19th century mathematicians who cultivated the subject nevertheless. Major contributors to the continuing development and application of probability theory during the 19th and 20th centuries included Cournot, Mill, Venn, Gauss, Poisson, Chebyshev, Markov, Bertrand, Poincaré, Hilbert, Khinchine, Kolmogorov, Reichenbach, and Keynes.

Practical Applications

Although much of their work involved games of chance and the solutions of abstract problems, many of the contributors to the development of theory found ways to apply their work to socially significant problems. Some, including Daniel Bernoulli and Jean le Rond D'Alembert, applied probability theory to the question of the advisability of the use of smallpox vaccination; Bernoulli, Euler, and others applied it to the computation of annuities; some calculated rates for insurance that took account of the probabilities of specific insurable events; several used the theory to argue the irrationality of participating in lotteries; Laplace applied it to problems in astronomy, to voting, and—especially with his invention of the method of least squares—to the prediction of error in many contexts.

As in mathematics more generally, problems that might appear to be solved simply for the intellectual challenge they posed were later turned to some practical purpose. Daniel Bernoulli, for example, solved the following problem: "In a bag are $2n$ cards; two of them are marked 1, two of them are marked 2, two of them are marked 3, ... and so on. We draw out m cards; required the probable number of *pairs* which remain in the bag" (Todhunter, 1865/2001, p. 229). Subsequently he found the solution to the problem useful in calculating tables of remaining life expectancies of married couples (as couples), letting the couples be represented by the pairs of cards in the problem.

People who wished to apply notions of probability or statistics to practical problems did not feel compelled to wait until a firm theoretical foundation had been built before doing so. Indeed, probability theory is an existence proof, *par excellence,* of the principle that the usefulness of mathematics often outruns its theoretical justification. It is one of several examples that could be given of an area of mathematics whose rigorous development was spurred by interest in practical problems to which it could be applied and, as such, it gives the lie to the idea that the best mathematics invariably comes from interests other than practical ones. As Daston (1987a) puts it, "The doctrine of aleatory contracts stimulated the development of a mathematics of chance and provided the fledgling theory of probability with both problems and a conceptual framework within which to solve them" (p. 254). Today probability theory is among the most extensively used areas of mathematics, despite continuing debates about the meanings of its foundational concepts; Kac (1964) refers to it as "a cornerstone of all the sciences" (p. 95). The ways in which this area of mathematics has been applied have had a profound impact on how we view ourselves and the world.

The importance of probability and statistics in the physical sciences was established by their application to a variety of problems, first to astronomy notably by Laplace (Stigler, 1987), and later to the kinetic theory of gases by Maxwell, Boltzmann, and Gibbs, to the analysis of Brownian motion by Einstein and Smoluchowski, and to the theory of radioactive decay by Rutherford. Pais (1986) identifies Rutherford's discovery of a half-life and Planck's development of the theory of the quantum at about the same time as the two events that marked the end of the era of classical physics and necessitated a fundamental revision of the concept of causality. That atoms decay spontaneously and that the probability that a given atom will decay instantaneously is constant over time are sufficiently well established in modern physics to be viewed as facts and are invoked to explain a wide range of other phenomena. Why atoms have this probabilistic property—why one cannot tell that a *particular* atom is more or less likely than others of the same type to decay in a given time—remains a mystery. The theory of the quantum has made probability central to particle physics in many ways; indeed, one may say that the theory of probability is fundamental to quantum physics.

More generally, although the term *revolution* is undoubtedly overworked, it seems appropriate to apply it to the change in thinking effected by probability, chance, and closely related concepts during the 19th and early 20th centuries. As I. B. Cohen (1987) puts it:

> The physical, biological, and social scientists of the twentieth century are almost universally aware that the establishment of a statistically based physics (radioactivity and quantum physics), biology (especially genetics), evaluation of experimental data, and social science have constituted so sharp a break with the past that no term of lesser magnitude than revolution should be used to characterize this mutation. (p. 34)

Hacking (1987b) highlights the transformation in thinking that occurred by pointing out that for Hume chance was "nothing real," whereas for von Neumann, it was perhaps the only reality. Regarding whether *revolutionary* is the right term to apply to the effect that notions of probability and chance have had on scientific thinking, Hacking concludes this way: "What is clear, beyond all scholasticism, is this. The taming of chance and the erosion of determinism constitute one of the most revolutionary changes in the history of the human mind. I use the word 'revolutionary' not as a scholar but as a speaker of common English. If *that* change is not revolutionary, nothing is" (p. 54). Krüger (1987b) argues that "if probabilism means to accord probability an explanatory function, it could not be realized short of a revolution in thought, a rethinking of the relationship of human reason and factual contingency" (p. 84).

Surprisingly, the revolution in thinking did not extend initially to economics. Ménard (1987) points out that leading 19th-century economic theorists made little or no use of probability in their writings: "Cournot's leading book, *Researches into the Mathematical Principles of the Theory of Wealth* [published in 1838], did not use probability at all. On the contrary, the first rigorous model in economic theory was a rejection of the idea that probability could be a useful tool at the core of economic analysis" (p. 140). This is the more surprising in view of the fact that Cournot, the author of *Exposition de la Théorie des Chances et des Probabilités (Exposition of the Theory of Chance and Probability),* which was published in 1843, was an expert on probability theory and a contributor to its development. As to why the impact of probability on economic thinking was delayed relative to its impact in other areas, one hypothesis is that certain preconditions had to be met before probabilistic ideas could be applied to advantage in this context, and these were not met until well into the 20th century (Horváth, 1987).

Ménard (1987) contends that the probabilistic revolution in economics has yet to happen. M. S. Morgan (1987), in contrast, argues that, although the probabilistic revolution in economics was surprisingly delayed, it did occur in econometrics in 1944 with the publication of Trygve Haavelmo's "The Probability Approach in Econometrics": "Economists' perception of the role of probability theory in the 1910s to the 1930s was that it had a very narrow domain of application, which extended neither to the treatment of economic data nor to the activity of measuring and uncovering economic laws; still less was probability seen as an element in theory itself" (p. 172). Morgan points out that during this period, statistical methods (e.g., least squares) that did not need to be justified by probability theory were widely used by economists, but by the end of the 1930s, things had begun to change and probabilistic concepts had started to find their way into the theoretical models economists were developing; in the mid-1940s, Haavelmo made a strong, and apparently successful, case for continuing and expanding this trend.

WHAT IS PROBABILITY?

> The vision and the intuition of the fair coin are fogged over and the road to axiomatization is beset with pitfalls. So also are the philosophical and psychological bases of probability. (P. J. Davis & Hersh, 1981, p. 165)

> Despite the many attempts no agreement as to what probability is appears to be in sight. (Macdonald, 1986, p. 16)

> The abstract noun "probability"—despite what we learnt at our kindergartens about nouns being words that stand for things—not merely has no tangible counterpart, referent, *designatum* or what you will, not merely does not name a thing of whatever kind, but is a word of such a type that it is nonsense even to talk about it as denoting, standing for, or naming anything. (Toulmin, 1958, p. 65)

The concept of probability has been around in something close to its current form for more than 300 years and has proved to be extremely useful. The theory of probability is now an established area of mathematics with a well-stocked store of axioms, theorems, corollaries, lemmas, and the rest. What exactly "probability" means, or should mean, however, is far from established. Is there such a thing as objective probability as distinct from a state of mind? Is probability a property of a proposition or of an event? Is the concept of chance incompatible with that of determinism? How does one determine the probability of a possible event? These and many similar questions have been the subjects of debate and discussion that began in the beginning and continue to this day.

One point on which there appears to be little debate is that whatever probability is it is closely related to the concept of chance. Uspensky (1937) put it this way: "It is always difficult to describe with adequate conciseness and clarity the object of any particular science; its methods, problems, and results are revealed only gradually. But if one must define the scope of the theory of probability the answer may be this: The theory of probability is a branch of applied mathematics dealing with the effects of *chance*" (p. 1). Certainly in what follows, we shall have many occasions to refer to chance, and the closely associated concept of randomness; it would not be possible to discuss probability without use of them.

As to what constitutes chance, Laplace (1814/1951) offered the following definition:

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible. (p. 7)

Laplace argued that when the probability is unknown, all possible values of probability between 0 and 1 should be considered equally likely. (This view has been controversial [Hacking, 1975; Polya, 1954b].) Together, Upsensky's definition of probability and Laplace's definition of chance make a fairly tight circle; the idea of equal possibility or likelihood keeps popping up in discussions of both of these concepts. Laplace did, however, make provision for situations in which all the cases are not equally possible. In such cases, one should "determine first their respective possibilities, whose exact appreciation is one of the most delicate points of the theory of chance. Then the probability will be the sum of the possibilities of each favorable case" (p. 11). (Laplace often used *possibility* in much the way we today use *probability.*)

Peirce (date unknown/1956) referred to the theory of probability as "simply the science of logic quantitatively treated" (p. 1334): "The general problem of probabilities is, from a given state of facts, to determine the numerical probability of a possible fact. This is the same as to inquire how much the given facts are worth, considered as evidence to prove the possible fact. Thus the problem of probability is simply the general problem of logic" (p. 1334). Despite his use of *simply* in these comments, Peirce saw probability as a controversial and not well understood subject. "It may be doubted," he claimed, "if there is a single extensive treatise on probabilities in existence which does not contain solutions absolutely indefensible" (p. 1334). This state of affairs he attributed in part to the lack of a regular method for calculating probabilities and in part to the disputed nature of the fundamental principles of its calculus.

Probabilistic reasoning is an example, in Peirce's (date unknown/1956) view, of explicative, analytic, or deductive reasoning, which he distinguished from amplificative, synthetic, or, "loosely speaking," inductive reasoning. He interpreted probability frequentistically and defined the probability of a mode of argument "as the proportion of cases in which it carries truth with it" (p. 1336). The probability of the conclusion B given the premise A, for example, was seen as the ratio of the number of times in which both A and B are true to the total number of times in which A is true independently of the truth of B. Peirce also recognized the distinction between probability "as a matter of fact," which is calculated as a ratio of two sets of events, one of which is a subset of the other, and probability as "the degree of belief which ought to attach to a proposition" (p. 1342), although he claimed that most writers had mixed the two conceptions together. The latter view, "though answering well enough in some cases" (p. 1346), he considered to be generally inadequate.

In contrast to the view that probabilistic reasoning is a form of deduction is a close association of it with induction. "The two topics, induction and probability, are nowadays so closely linked in philosophers' minds, that no explanation is felt to be needed why a book, for example, which is about induction should also be about probability" (Stove, 1986, p. 113). Stove argues that probability and induction are so closely linked in philosophers' minds that they "are often inextricably confused" (p. 113). Perhaps the point on which one would be most likely to get consensus among philosophers, mathematicians, and others who make use of the concepts of probability and chance is that it is difficult indeed to find definitions on which one can expect general agreement. As Carnap (1953) put it: "Practically everyone will say that probability as used in science has only one meaning, but when you ask what that meaning is, you will get different answers" (p. 128).

States of Mind or Something More?

> Any probability should in principle be indexed with the name of the person, or people, whose opinion it describes. (Edwards, Lindman, & Savage, 1963, p. 197)

A question about probability that has been around almost as long as the concept and that persists to this day is whether statements such as "The probability that the next toss of the die will come up 4 is 1/6" and "The chance of rain tomorrow is about .8" should be viewed as anything more than reflections of states of mind—expressions of expectations, of degrees of knowledge or ignorance, about the outcomes of uncertain events, or of levels of confidence in the truth of specified assertions.

According to one view, the idea of chance is necessitated only by the limitations of human knowledge. This view was held by many of the major contributors to probability theory, including Laplace and de Morgan. According to it, all apparently probabilistic phenomena are, in fact, deterministic. They are unpredictable because of our ignorance of the cause–effect relationships involved, or because, even if we understand the relationships in principle, the situations are too complex to be tractable computationally.

Hacking (1990) notes that the association of chance and probability with lack of knowledge prevailed during what we now refer to as "the Age of Reason." There is, in his characterization of the view, a hint of something bordering on the disreputable; the need to invoke such concepts appears to have been seen as an embarrassment to the human intellect:

> Chance, superstition, vulgarity, unreason were of one piece. The rational man, averting his eyes from such things, could cover chaos with a veil of inexorable laws. The world, it was said, might often look haphazard, but only because we do not know the inevitable workings of its inner springs. As for probabilities—whose mathematics was called the doctrine of chances—they were merely the defective but necessary tools of people who know too little. (p. 1)

An alternative view is illustrated by Poincaré's (1913/1956) position that "chance is something other than the name we give our ignorance" (p. 1381). We must, he suggests, distinguish between two types of phenomena the causes of which are unknown to us: those "fortuitous phenomena about which the calculus of probabilities will provisionally give information" and "those which are not fortuitous and of which we can say nothing so long as we shall not have determined the laws governing them" (p. 1381). Poincaré recognized that very slight causes can have very large effects and believed that often when the causes are sufficiently slight to escape our notice we (mistakenly) say the effects are due to chance. This idea has found more recent expression in the concept of chaos and complexity theory, according to which some events that appear to be random are really deterministic but too complex to be predictable, except probabilistically; the occurrence of such events does not rule out, of course, the existence of other events that are unpredictable, except probabilistically, in principle. Statements of probability can be seen as based on lack of knowledge in both cases, but knowledge that is available in principle in one case and not in the other.

The theory of probability was affected by the 19th-century development of statistics as a discipline and a frequency interpretation was worked out by several people in the 1830s and early 1840s. According to the statistical, or frequency, view, which was held by John Stuart Mill, R. L. Ellis, Antoine Augustin Courant, and J. F. Fries, and worked out in detail by John Venn in the 1860s, "statements of probability are properly understood as being about the frequency of like events in the long run, and not as measures of expectation concerning the outcome of some particular uncertain event" (Gigerenzer et al., 1989, p. 45). Gigerenzer (1994) argues that failure to distinguish between frequencies and single-event probabilities has led to much confusion in the psychological literature on probabilistic thinking.

But even on a strict frequentistic interpretation of probability, it is not clear that one can expect unanimity with respect to what it means in all contexts. Consider, for example, Toulmin's (1958) claim that "no one person is permitted, in one and the same breath, to call the same thing both improbable and true ... to do this is to take away with one hand what is given with the other" (p. 54). According to Toulmin's position, it is inappropriate to speak of an event that one knows to have occurred as improbable; it once may have been appropriately considered improbable, but not after it is known to be true. To say that something that is known to have happened *sounds* improbable, or *seems* improbable, or is *believed* by someone else to be improbable is permitted, but to say that it *is* improbable is not. Moreover, if an event has occurred, someone who does not realize that it has occurred and who believes its occurrence to be improbable is mistaken in that belief. Similarly, if I believe it to be probable that such and such an event will occur at such and such a time, its failure to occur at the specified time shows my belief to have been mistaken. Toulmin is careful to distinguish between a claim that was improper when it was made and one that subsequently turned out to have been mistaken; the fact that a claim proved to be mistaken does not demonstrate that it was improper as judged in light of the evidence on which it was based.

This view seems to me to carry the implication that *all* probabilistic claims are mistaken. All events that are assigned a probability eventually either occur or do not, which is to say they all prove to have a probability of either 1 or 0. So, according to this view, if I say the probability that the next toss of a die will come up 3 is 1/6, that is a proper belief for me to hold, but I should realize that it is a mistaken one; as soon as the die is tossed, I will discover that the probability was not 1/6 at all but either 1 (if it comes up 3) or 0 (if it does not). This strikes me as a tenable view, but not a particularly helpful or attractive one.

Some writers make a distinction between a probability and a degree of confidence. Suppose a fair coin has already been tossed, but I do not know how it has landed. Some, who might allow that before the coin was tossed the probability that a toss would come up head was .5, would argue that the probability no longer applies after the toss; after the fact, they would claim, the appropriate thing for me to say is that my degree of confidence that it has landed head is .5 (Reichardt & Gollob, 1997).

One Meaning or Many?

Citing B. J. Shapiro (1983) and Daston (1988), Gigerenzer (1994) notes that many connotations of probability were promoted at different times during the 18th and 19th centuries, including "physical symmetry (e.g., the physical construction of dice, now called 'propensity'); frequency (e.g., how many people of a given age died annually); strength of argument (e.g., evidence for or against a judicial verdict); intensity of belief (e.g., the firmness of a judge's conviction in the guilt of the accused); verisimilitude; and epistemological modesty, among others" (p. 134). A remarkable aspect of the use of the concept by Enlightenment probabilists, Gigerenzer suggests, is the ease with which they slid from one meaning of the term to another.

Writing in the 18th century, D'Alembert suggested the need for a distinction between what is *metaphysically possible* and what is *physically possible*. Anything is metaphysically possible that is not ruled out as a conceptual absurdity; something is physically possible only if it could actually occur: "It is *metaphysically* possible to throw two sixes with two dice a hundred times running; but it is *physically* impossible, because it never has happened and never will happen" (Todhunter, 1865/2001, p. 262).

The considerable attention given to probability during the 20th century did not result in a convergence on a definition to which all theorists would subscribe. Not only do different writers continue to conceptualize probability in different ways, many explicitly recognize two or more types of probability and discuss the circumstances under which each meaning is to be used. W. Weaver (1950), for example, makes a distinction between *mathematical probability* and *statistical probability*. For mathematical probability, "equally likely cases" is an undefined concept, playing much the same role in probability theory that points and lines play in geometry. For statistical probability, the likelihood of cases is determined empirically; the likelihood that a randomly selected child will be a girl is the percentage of all children that are girls. Margolis (1987) distinguishes between probability$_g$, or probability in the gambling sense, and probability$_b$, which connotes believability or plausibility. He argues that sometimes participants in experiments on probabilistic reasoning answer questions on the basis of probability$_b$ when the experimenter intended that they answer them on the basis of probability$_g$.

Nagel (1936/1956) distinguishes three interpretations of probability: the view formulated by de Morgan (and held more recently by Harold Jeffreys [1934, 1939] and L. J. Savage [1954/1972, 1962] among others), which sees probability as a state of mind or the strength of belief in the truth of a proposition; the view, associated with John Maynard Keynes (1921), that probability is an unanalyzable but intuitively understandable logical relation between propositions; and the view that equates probability with relative frequency in the same sense as noted earlier. Nagel argues that probability can have different meanings in different contexts. He sees the frequency interpretation (when properly qualified) as the most satisfactory one in everyday discourse as well as in applied statistics and measurement, and in many branches of the theoretical sciences. He sees it also as appropriate for interpreting some statements about the probability of hypotheses, but does not think it the appropriate interpretation in statements about the probability of complicated scientific theories.

Keynes (1921/1956) treats the concept of probability almost as an undefinable primitive:

> No knowledge of probabilities, less in degree than certainty, helps us to know what conclusions are true, and ... there is no direct relation between the truth of a proposition and its probability. Probability begins and ends with probability. That a scientific investigation pursued on account of its probability will generally lead to truth, rather than falsehood, is at the best only probable. The proposition that a course of action guided by the most probable considerations will generally lead to success, is not certainly true and has nothing to recommend it but its probability. The importance of probability can only be derived from the judgment that it is *rational* to be guided by it in action; and a practical dependence on it can only be justified by a judgment that in action we *ought* to act to take some account of it. (p. 1373)

Keynes expects that people with the same education would tend to agree regarding what the probability of a specified uncertain event would be.

Ayer (1965) distinguishes three types of judgments of probability: judgments of a priori probability, statistical judgments, and judgments of credibility. As illustrations of the three, he gives a statement of the likelihood of tossing double six with a pair of true dice, the observation that any given unborn child is slightly more likely to be a boy than a girl, and the claim (as of 1965) that there is little chance that Britain will join the Common Market. Ayer distinguishes also five different types of events that are described as happening by chance: (a) a member of a series that conforms with the a priori calculcus of chances (theory of probability), (b) a deviation from an established frequency, (c) an event that was not intended by the agent that caused it, (d) concurrent (coincidental) events for which there is no ready causal explanation, and (e) specific outcomes of statistically determined processes (on average, one of six tosses of a die will show a three; *which* tosses show a three is a matter of chance in this sense).

L. J. Cohen (1982) identifies five different interpretations of probability: "We can interpret the probability-function that is regulated by the classical calculus of chance either as a ratio of *a priori* chances, or as a relative frequency of empirically given outcomes, or as a causal propensity, or as a logical relation, or as a measure of actual or appropriate belief" (p. 253). He does not contend that one of these interpretations is right and the others wrong: "Just as we can usefully measure quantities of apples in at least three different ways (by number, by weight or by volume), so too, if we desire, we can measure probabilities in several different ways according to the purpose in hand" (p. 253).

Continuing the escalation, Good (1983b) explicitly distinguishes seven kinds of probability: tautological (mathematical), physical, and five types of intuitive (logical, subjective, multisubjective, evolving, and psychological). He defines subjective probability as "psychological probability to which some canons of consistency have been applied" and logical probability as "the subjective probability in the mind of a hypothetical perfectly rational man" (p. 122). Logical probability is sometimes referred to as credibility.

In sum, two points are clear: First, there is not a consensus among people who use probability theory about what probability means, and second, it is possible to distinguish several different connotations that have been given to the term. A comment by Toulmin (1958) describes the situation well:

> The attempt to find some "thing," in terms of which we can analyze the solitary word "probability" and which all probability-statements whatever can be thought of as really being about, turns out therefore to be a mistake.... To say that a statement is a probability-statement is not to imply that there is some one thing which it can be said to be about or express. There is no single answer to the questions, "What do probability-statements express? What are they about?" Some express one thing: some another. Some are about to-morrow's weather: some about my expectation of life. If we insist on a unique answer, we do so at our own risk. (p. 70)

G. Shafer and Tversky (1985) express a similar view in characterizing theories of subjective probability as "formal languages for analyzing evidence and expressing degrees of belief" (p. 309), and arguing that, because the picture of chance can be related to practical problems, the probability languages that can be constructed can differ with respect to both semantics and syntax. It need not be assumed that all of these languages have equal normative claims, they contend, and there is "the possibility that no single language has a preemptively normative status" (p. 315). Shafer and Tversky discuss specifically the language of Bayesian probability and that of belief functions. making finer distinctions within this broad one. (Bayesian probability is discussed in chap. 4.) Within Bayesian probability, for example, they distinguish the following three semantics: frequency semantics (which "compares our evidence to the scale of chances by asking how often, in situations like the one in hand, the truth would turn out in various ways"), propensity semantics (which "makes a comparison by first interpreting the evidence in terms of a causal model and then asking about the model's propensity to produce various results"), and betting semantics (which "makes the comparison by assessing our willingness to bet in light of evidence: at what odds is our attitude towards a given bet most like our attitude towards a fair bet in a game of chance?") (p. 316).

Much more could be said by way of illustrating that there is not general agreement among people who think about such things regarding what probability and closely associated concepts mean, but enough has been said to make the point. From a pragmatic perspective, it is not necessary that there be a single agreed-upon definition of probability for the concept and the mathematics that has been developed around it to be of use, and indeed the mathematics of probability theory is used to great advantage by people holding very different views on the question of what probability "really means" as well as by those who hold none at all. This fact makes the question no less interesting to those who like to think about such matters, but it may be reassuring to those who wish to use the concept and associated mathematics for practical purposes, without concerning themselves much with philosophical debates about meaning.

Probability theory provides us with a tool for making useful predictions at certain levels of observation. It permits us to deal with the behavior of gases or human populations, for example, by treating probabilistically the motions of the molecules of gas or the behavior of the individuals that comprise a group; and this is a very great help, because even if the motion of every single molecule or the behavior of every individual is assumed to be deterministic and to obey known laws, there are simply too many of them to allow the inferring of the behavior of gases or populations from calculations of their individual trajectories or actions.

Probability and Determinism

> The historical relationships of probability theory and statistics with determinism and indeterminism ... defy any simple generalization. Probability and statistics have served both masters at one or another point in their history, depending on which interpretation of probability was then in the ascendant. (Gigerenzer et al., 1989, p. 276)

> The disavowal of determinism was, ironically but essentially, a development of the statistical tradition. It represents one of the most interesting outcomes of the introduction of statistical thinking to science, social thought, and philosophy in the nineteenth century. (Porter, 1986, p. 227)

> There *are* people who say that it is merely extremely probable that water over a fire will boil and not freeze, and that therefore strictly speaking what we consider impossible is only improbable. What difference does this make in their lives? Isn't it just that they talk rather more about certain things than the rest of us? (Wittgenstein, 1953/1972, p. 43)

Probability and determinism are often viewed as antithetical ideas. The behavior of deterministic systems is considered to be lawful and, at least in principle, predictable. That of probabilistic systems is seen to be erratic and inherently unpredictable in detail. In fact, completely deterministic systems can produce unpredictable behavior. Mathematical chaos, complexity theory, and other areas of mathematics that focus on nonlinear systems present many examples of unpredictably complex—some would say truly random—behavior produced by systems whose operation is governed by relatively simple deterministic rules. Some observers believe this fact reveals a deep truth about the universe and our ability to know it: "The conclusion must be that even if the universe behaves like a machine in the strict mathematical sense, it can still happen that genuinely new and in-principle unpredictable phenomena occur" (Davies, 1988, p. 55).

Not only can deterministic systems be unpredictable, but the behavior of probabilistic systems can be highly predictable in the aggregate; at the appropriate level of description, probabilistic phenomena are remarkably lawful and predictable, and this is a primary reason why probability theory has been of such great interest and of use in so many contexts. It would seem that "we can't get away from determinism. Chase it out the door, by postulating total incoherence, and it comes back through the window, in the guise of statistical laws" (Ekeland, 1993, p. 50).

The relationship between probability and determinism is not easy to describe in a few words, in part because both terms are used to represent a variety of concepts. We have already noted several connotations that have been given to probability. Determinism has been conceived in several ways as well. Gigerenzer et al. (1989) distinguish five—metaphysical, epistemological, scientific, methodological, and effective—and argue that whereas "the empire of chance" has shaken scientific determinism, it has left the other types intact.

Some students of chance have seen statistical regularity as a form of determinism. De Moivre (1756/1962), for example, expressed that view this way: "If from numberless observations we find the ratio of events to converge to a determinate quantity, as to the ratio of P to Q, then we conclude that this ratio expresses the determinate law according to which the event is to happen" (p. 264). Buckle (1857–1861) saw statistical regularities as evidence of a lawfulness that could be countermanded neither by chance nor by human will. Davies (1992) captures something of this idea in declaring a difference between stochasticity and anarchy.

Heated and extensive debate has occurred also as to whether the phenomenon of statistical regularity should be seen as evidence against human freedom. No one denies the reality of the phenomenon; the debate has centered on the question of whether aggregate regularity is consistent with freedom at the level of the individual. Like most such debates, this one has had no clear winner, or at least no losers willing to acknowledge themselves as such.

A particularly interesting aspect of the debate, because it tells us something about human reasoning, is the contrast between attitudes prevailing in the early part of the 19th century and those more common a century later: "In the 1930s, the conviction that the laws of nature are probabilistic was thought to make the world safe for freedom. The incoherence went in the opposite direction in the 1830s: if there were statistical laws of crime and suicide, then criminals could not help themselves. In 1930, probability made room for free will; in 1830, it precluded it" (Hacking, 1990, p. 136). Hacking goes on to note, however, that doubts about whether the fact of statistical regularity really does support the idea of free will persist: "The cool-headed analytic view says that a statistical law may apply to a population, but members of the population remain free to do as they please. The law applies only to the ensemble of individuals.... Despite this glib and comfortable opinion, we have not made our peace with statistical laws about people. They jostle far too roughly with our ideas about personal responsibility" (p. 117).

Given the course that physics took during the 20th century, the history of the development of probability theory and statistics now seems a little ironic. Quetelet and others wanted to quantify social and psychological phenomena in the fashion of the physical sciences, in order to give the former the certainty enjoyed by the latter, and the application of statistics to social and psychological phenomena was central to this effort. Porter (1986) summarizes the attitudes among social scientists regarding the application of statistics to their field this way:

> The evident success of statistics as an approach to social science was not interpreted by contemporaries as vindication of a metaphysic which regarded the laws governing certain domains as only probable. On the contrary, statistical laws were deliberately formulated to extend the certainty of sciences like astronomy and mechanics to knowledge of phenomena which hitherto had resisted exact scientific investigation. (p. 69)

The irony is that an area of mathematics that was seen to be the avenue to certainty regarding social and psychological phenomena subsequently became instrumental in relieving us of much of the certainty with which we had viewed the physical world before the appearance of quantum mechanics on the scene. The acceptance of indeterminism has been referred to as "one of the most striking changes of modern scientific thought" (Porter, 1986, p. 149), and the discovery that the physical world is not deterministic as "the most decisive conceptual event of twentieth century physics" (Hacking, 1990, p. 1).

In quantum mechanics, the probabilistic nature of the properties of subatomic particles is considered to be not a matter of the limitations of our knowledge, but inherent to the particles themselves. Our inability to determine both the position and momentum of a particle to a high degree of accuracy is not just a consequence of the crudeness of our measuring techniques, and it is not even only a result of the fact that the act of determining one of these properties has an effect on the other property; according to some interpretations of the theory, a particle does not *have* a precise location and a precise momentum at the same time.

The emergence of probability theory and statistics as mathematical disciplines was not the only—perhaps not even the primary—cause of the demise of determinism in physics. This was forced by the results of experiments that could not be accounted for in classical deterministic terms. The fact that by the time these experimental results were obtained, probability theory and statistics were well established meant that scientists had the concepts and tools they needed to accommodate the new results. One wonders what course physics might have taken had these concepts and tools not been at hand.

Surprisingly, acceptance of the idea that the universe is nondeterministic at the quantum level has had the effect of greatly increasing physicists' ability to predict phenomena at higher levels of organization. Discovery of the indeterminate character of the building blocks of nature has somehow made their behavior in the aggregate more understandable and susceptible to modification and control. In short, indeterminism at the level of the individual particle does not seem to rule out the possibility of determinism at the level of their aggregate behavior. This observation applies equally to the behavior of inanimate physical systems and to that of individual people and large groups. Koestler (1972) speaks of the paradox of the lawfulness of chance: "The paradox consists, loosely speaking, in the fact that probability theory is able to predict with uncanny precision the overall outcome of processes made up out of a large number of individual happenings, each of which in itself is unpredictable. In other words, we observe a large number of uncertainties producing a certainty, a large number of chance events creating a lawful outcome" (p. 25).

Probability as a Mathematical Discipline

Neither philosophers nor mathematicians have been able to converge on a consensus as to what probability "really is," and it seems unlikely that such a consensus will emerge any time soon. This conceptual impasse appears not to have impeded the development of probability theory as a mathematical discipline, however, at all. William Feller, whose 1957 text on probability theory has long been recognized as a classic, declined to define probability: "We shall no more attempt to explain the 'true meaning' of probability than the modern physicist dwells on the 'real meaning' of mass and energy or the geometer discusses the nature of a point. Instead, we shall prove theorems and show how they are applied" (p. 3). As a calculus, the theory of probability works just fine without the benefit of such a definition, and its usefulness in countless contexts has been established beyond doubt.

One way to characterize probability theory as a mathematical discipline is as a set of rules for "calculating the probabilities of complex events consisting of collections of 'elementary' events whose probabilities are known or postulated" (Kac, 1964, p. 98). To apply the calculus of probability it is not necessary to know what it means, at a deep level, to say that the probabilities of specific elementary events are thus and so or to know how those probabilities were established; it is necessary only that the probabilities be specified numerically and that the values used conform to certain constraints (e.g., be positive values between 0 and 1 inclusive).

As a mathematical discipline, the calculus of probability functions like any other branch of mathematics. Its corpus is a set of theorems that are deducible from a few axioms and undefined primitives. And as is the case with other branches of mathematics, the entities with which it deals—the symbols it manipulates—can be viewed as abstractly as one might wish; they need bear no relationship to any aspects of the physical world. As it happens, much of the interest in probability theory—as in many other areas of mathematics—derives from the fact that when its symbols are taken to represent specific aspects of the physical world, the inferences that can be drawn by application of the symbol-manipulating rules of the calculus prove to be remarkably close to what can be determined by observation.

This correspondence is not a property of the mathematics itself, but a useful bonus. In the probability calculus, the joint probability of two independent events is the product of the probabilities of the individual events as a matter of definition. The definition stands whether or not there are such things as independent events in the physical world. No conclusions about the world can be drawn from the calculus without the involvement of some assumptions that do not come from the calculus itself. I may say that *if* real-world variables A (say the toss of a three with a specific die) and B (say the toss of a two with a different specific die) are independent and each is probabilistic, then I expect that the probability of their joint occurrence is equal to the product of their individual probabilities of occurrence. I would say this on the strength of my belief that the "laws of probability," by which I mean the theorems of probability theory, are descriptive of certain types of events in the physical world; but the belief is neither part of the calculus nor essential to it.

As Ayer (1965) puts it, "In themselves the propositions of the calculus are mathematical truisms. What we can learn from them is that if we assume that certain ratios hold with respect to the distribution of some property, then we are committed to the conclusion that certain other ratios hold as well" (p. 46). In the absence of an assumption about the applicability of probability theory to the physical world, we are committed to the indicated conclusion only in the abstract realm of the calculus itself. We make the assumption with some confidence because we have considerable evidence that it is true, but that evidence comes not from the calculus but from observation of real-world events. This dictates caution in applying probability theory to the description of real-world events. Rozeboom (1997) makes the point explicitly: "Until such time as the foundations of applied probability theory become rather less mysterious, we should exercise prudence in how unconditionally we trust the statistical models through which we filter our empirical research findings" (p. 366). This is not to contend that such models should not be used, but only to note the appropriateness of some care and tentativeness in the interpretation of the results of their use.

JUDGMENTAL VERSUS STATISTICAL PROBABILITY

> Probability has two aspects. It is connected with the degree of belief warranted by evidence, and it is connected with the tendency, displayed by some chance devices, to produce stable relative frequencies. (Hacking, 1975, p. 1)

> What guarantee have we for assuming that limits of relative frequencies bear any significant relation to finite relative frequencies, which are, after all, the only things we can observe? (Von Plato, 1987, p. 393)

Although several possible meanings of probability have been articulated by people who have thought about the subject, the distinction that has been most widely recognized is the dichotomous one between judgmental (subjective, intuitive, epistemological, inductive, epistemic) and statistical (objective, physical, aleatory) probability. Some variant of this distinction is recognized by most modern writers. Polya (1954a, p. 116) makes this distinction, for example, when he points out that the calculus of probability can be used both to systematize the rules of plausible inference and to describe random mass phenomena. In the first case probability refers to credibility or degree of reasonable belief and in the second to long-range relative frequency. He considers both uses legitimate, but warns against confusing the two.

Kamlah (1987) also makes this distinction in discussing the history of the concept of probability: "[T]he more probabilistic laws, such as Mendel's laws or the statistical laws of Brownian motion or radioactive decay, became known in natural science, the more it became clear that physical probabilities are physical quantities, while subjective probabilities are at best methodological tools—if they can be justified at all" (p. 113). Kamlah's claim that physical connotation of probability gained popularity over time suggests that subjectivism was stronger among the original contributors to the development of probability theory. Daston (1987b) makes this explicit: "The apparent miscellany of applications to gambling, insurance, astronomy, medicine, reliability of testimony, accuracy of tribunal judgments, economic theory of value, and reasoning from known effects to unknown causes were in fact joined by a single thread: all problems were posed in terms of reasonable belief and action based upon that belief" (p. 297). This perspective is sometimes referred to as Laplacian inasmuch as his concept was strongly subjectivist. Sometimes the distinction between subjective and objective probability was made on the basis of whether the event(s) in question occurred only once or repeatedly; probability was considered subjective in the former case and objective in the latter (Jorland, 1987).

Carnap (1950/1962, 1953) distinguishes between probability as a logical concept—inductive probability—which refers to the amount of evidentiary support for some hypothesis or claim, and probability as an empirical concept—statistical probability—which connotes the frequency of an event of interest relative to the frequency of all events in the same class (e.g., the frequency of the toss of a three with a die relative to the total number of tosses of the die). He sees both types of probability as indispensable to science, but argues the importance of maintaining a distinction between them. Salmon (1974) too distinguishes between probability as degree of rational belief and probability as relative frequency. Hacking (1975) makes a similar distinction, referring to probability that connotes degree of belief as epistemological probability and to that which connotes stable frequencies as aleatory probability.

Although, as already noted, Feller (1957) declines to define probability in his textbook treatment of probability theory, he does distinguish between intuitive or judgmental probability and physical or statistical probability. The former has to do with beliefs or states of mind, the latter with possible outcomes of a conceptual experiment. Fundamental to physical or statistical probability is the concept of a "sample space," the points of which represent the full set of possible outcomes of the conceptual experiment. The sample space and its points are the primitive undefined notions of the theory and occupy within it the same status as do the notions of point and straight line in Euclidean geometry.

A distinction of this sort was made explicitly at least as early as 1837. Both Poisson (1837) and Cournot (1843) use the word *probabilité* to represent reason for belief that an event has occurred or will occur and *chance* to indicate an objective property of an event, namely its propensity or "facility" to occur. Hacking (1975) argues that the dual character of probability can be seen even in the work of the several people who began to develop the theory independently around 1660, and in some precursor work, such as that represented by Cardano's book, which, as has already been noted, was written early in the 16th century but not published until 1663: "It is notable that the probability that emerged so suddenly is Janus-faced. On the one side it is statistical, concerning itself with stochastic laws of chance processes. On the other side it is epistemological, dedicated to assessing reasonable degrees of belief in propositions quite devoid of statistical background" (p. 12).

The correspondence between Pascal and Fermat regarding how to divide the stakes in a prematurely terminated game of chance concerns aleatory probability; Pascal's famous "wager" defending the reasonableness of belief in God concerns probability in the epistemological sense. Hacking (1990) contends, however, that whereas philosophically minded students take this and related distinctions seriously, the vast majority of users of probability theory for practical purposes do not, and some extremist proponents of one or another view often deny the distinction, recognizing as legitimate only the type of probability they espouse. Hacking notes that the two have been dominant at different times and argues that debate about which is correct is pointless. The change that time has effected in the connotation that probability has had in physics is seen in a comment by Beatty, Cartwright, Coleman, Gigerenzer, and M. S. Morgan (1987): "The probabilistic laws of classical statistical mechanics were supposed to be a function of human ignorance; those of quantum mechanics, to reflect the structure of nature" (p. 1).

A strong commitment to a frequentist interpretation of probability appears to have been the case among mid-19th-century philosophers and philosophically minded mathematicians with an empiricist bent, including Cournot, Ellis, Fries, Boole, Mill, and Venn (G. Shafer, 1993). Notable frequentists of the 20th century include Ronald A. Fischer and Richard von Mises. Heidelberger (1987) credits Gustav Fechner with doing the foundational work on which von Mises and other 20th-century frequentists built: "It was mainly Fechner's work that led to the decline of the classical Laplacian theory of probability and the rise of frequency theory in our century. Determinism in the nineteenth century allowed for indeterminism only as a result of human ignorance, while Fechner's theory led to the conception that probability theory is an empirical science of chance phenomena in nature" (p. 135). Basic to Fechner's theory were the concepts of a collective object and chance variation. Heidelberger describes Fechner's contribution in some detail.

G. Shafer (1993) argues that a resolution of the subjectivist–frequentist controversy, of sorts, was provided by Kolmogorov's axiomatization of probability theory early in the 20th century. The effect of this work was to differentiate clearly probability theory as a mathematical discipline from interpretations and practical applications thereof. Salsburg (2001) says of Kolmogorov's axiomatization that "it is taught today as the only way to view probability. It settles forever all questions about the validity of those [probability] calculations" (p. 144). As Shafer notes, however, Kolmogorov's resolution of the subjectivist–frequentist controversy has not been satisfactory to all. And Salsburg points out that Kolmogrov was himself keenly aware that it did not answer the question of what probability means in real life. This question—more philosophical or metaphysical than mathematical—has yet to be answered in a way that is recognized as compelling by all, or even most, of the people who have thought and written about it.

The distinction between judgmental and statistical probability seems to me to be an intuitively compelling one. The probability that the extinction of the dinosaurs was the result of a meteorite colliding with the earth seems different in kind from the probability of tossing five straight heads with a fair coin. Similarly, the probability that a woman will win a specific future U.S. presidential election seems a different sort of entity than the probability that a specified ticket holder will win a particular lottery. We speak of probabilities in all four cases, but in the second of each pair of examples we would expect to find a high degree of agreement among statisticians as to how to determine the probabilities and even regarding what they are, whereas for the first example in each pair we would not expect a strong consensus with respect to either what the probability is or how to determine it.

Although many writers recognize both judgmental and statistical interpretations of probability, many others recognize one or the other but not both. Some who see a probability statement as a description of a state of mind—an expression of a degree of uncertainty—hold that this is true whether one is talking about events, like births of male and female children, that happen all the time, about one-of-a-kind events, like the extinction of the dinosaurs, or about possible future events, like worldwide thermonuclear war, that have never happened in the past. Some theorists who hold a strictly frequentistic view of probability see all meaningful statements of probabilities either as assertions about the actual relative frequencies with which the events of interest have occurred in the past or, in the case of events that have occurred infrequently or not at all, as statements of the relative frequencies that *would* be observed if the events that made up the numerator and denominator of the ratio occurred often. The position has sometimes been argued by frequentists that application of probability to an individual unrepeatable event is not justified.

The distinction between two types of probability, one referred to as judgmental, subjective, intuitive, epistemological, inductive, or epistemic and the other as statistical, objective, physical, or aleatory, seems fundamental, but finer distinctions have also been made. Gigerenzer et al. (1989) note that the mathematicians who initially tried to measure probabilities came up with at least three methods: "equal possibilities based on physical symmetry; observed frequencies of events; and degrees of subjective certainty or belief" (p. 7). The second and third of these correspond directly to statistical and judgmental probability, respectively, as these terms are used here. The first category—equal possibilities based on physical symmetry—is generally considered an instance of statistical or physical probability; but there clearly is a difference between accepting 1/6 as the probability that a tossed die will land with three up on the strength of the physical symmetry of the die, and taking 18/35 to be the approximate probability that an unborn child will be male on the grounds that historical data show this to be the long-term relative frequency of male births.

What Is a Judgmental Probability?

Relative frequency is, for most observers, a noncontroversial concept, connoting or denoting simply ratios of numbers of events. To say that the relative frequency of heads in a large number of tosses of a coin was .507 is to say that, on the average, 507 of every 1,000 tosses were heads. No one quarrels about this use of terms. What is sometimes disputed is the use of empirically determined relative frequencies as the basis of making claims about probabilities. No one will question the appropriateness of saying that the relative frequency of male births is 18/35, or about .514, assuming that is what the records show; but some people will object to saying, on the strength of this fact, that the probability that the gender of some soon-to-be-born child (whose gender has not been determined by tests) will be a male about .514. On the other hand, there is the view that relative frequency is a more fundamental concept than probability and that people's beliefs about the latter derive largely from their observations of the former (Estes, 1976a, 1976b).

What exactly should a judgmental probability be taken to mean? When I say that I believe the probability of some possible future event to be X, what am I saying? One answer to this question invokes the notion of statistical probability in a theoretical way. It essentially says that *if* the appropriate experiment could be done, it would yield the event in question in the ratio X; the fact that the experiment cannot be done, if it cannot, does not preclude this interpretation. Another possible interpretation is that the statement of probability represents nothing more than the strength of a belief, expressed in such a way that 0 and 1 represent respectively the minimum and maximum strengths possible. One way of measuring strength of belief, and thus of subjective probability, is by determining one's willingness to bet on the event whose probability is in question, as proposed by Borel (Knobloch, 1987). There are other interpretations as well. Hacking (1975) distinguishes three different conceptions of probability among current subjectivists: the extreme form of subjectivism of L. J. Savage (1954/1972, 1962) and Bruno de Finetti (1976), the theories of inductive probability developed by J. M. Keynes (1921), and the subjective probability of quantum physicists. It suffices, for present purposes, to acknowledge these distinctions without pursuing them further here.

L. J. Cohen (1982), who stresses the ambiguity of judgmental probability, argues that the common view that a subjective probability assigned to an uncertain event is veridical to the extent that the assigned number reflects the objective probability of that event is correct in a trivial sense where the subjective probability of X means a person's estimate of the objective probability of X, but incorrect when the subjective probability of X means the strength of belief in X and the objective probability of X means the relative frequency of X. When people estimate the probability of a specific outcome, Cohen contends, they may take into account not only the relative frequency of that outcome but also any other relevant information they may have. One's strength of belief that a specified individual would survive to a given age could be informed, for example, not only by one's knowledge of the actuarial data that provide the relative frequency with which people in the demographic group to which the individual belongs survive to that age, but one's knowledge of the person's general state of health, survival-relevant habits, and so on.

L. J. Cohen (1982) notes further that a subjective probability can be taken to mean what one personally believes about some uncertain event, or it can be taken to mean what *anyone ought* to believe on the strength of the evidence. By the first, or descriptive, interpretation, different subjective probabilities held by different people are not necessarily contradictory, but by the second, or prescriptive, interpretation they are.

The Principle of Indifference

Carnap (1953) illustrates the distinction between subjective—his term is *inductive*—and statistical probability by noting that the "principle of indifference" can be interpreted in different ways when the former type of probability is in question. His illustration involves the drawing of four balls from an urn that contains blue and white balls in an unknown ratio. The question is how to assign a priori probabilities to the possible outcomes. The answer is obtained by listing all possible outcomes and applying the principle of indifference, according to which, in the absence of evidence to the contrary, all possible outcomes should be considered equally probable. Carnap argues that in this example, the principle can be applied in two ways, depending on what one considers an outcome to be.

One might consider an outcome to be a specific sequence of draws: blue, blue, white, blue; or one might consider an outcome to be the drawing of a particular ratio of blue to white balls: three blue, one white. Sixteen different outcomes are distinguishable in the former case, but only five in the latter. Applying the principle of indifference gives as the probability of each outcome 1/16 in the first case and 1/5 in the second. Unfortunately, when the 16 outcomes from the first analysis are mapped onto the 5 of the second, the probabilities assigned are seen to be inconsistent. There is only one sequence that yields four blues, for example, and this has probability 1/16 according to the first view and 1/5 according to the second; similarly, there are six sequences that contain two blues and two whites, so the probability of getting some one of these sequences, according to the first view is 3/8 (the sum of the probabilities of the individual sequences), whereas, according to the second application of the principle of indifference, the probability of getting two blues and two whites is 1/5.

I believe that most theorists would reject this illustration on the grounds that it is wrong to consider all possible blue-to-white ratios to be equally probable. Assuming that the drawing is with replacement, and letting the probability of drawing a blue ball be p and that of drawing a white one $1-p$, the probability of drawing k blues and $4-k$ whites in four draws would be

$$\binom{4}{k} p^k (1-p)^{4-k}.$$

In order for the probability of the drawing of k blues and 4-k white balls to be equal for all values of n from 0 to 4, it would have to be the case that

$$(1-p)^4 = 4p(1-p)^3 = 6p^2(1-p)^2 = 4p^3(1-p) = p^4,$$

and there is no value of p for which these equalities can all hold.

Nevertheless, the principle of indifference is a controversial one, and at least as sometimes expressed, can lead to contradictions. (See the discussion of Bertrand's paradox in chap. 6.) Carnap (1953) illustrates this by applying Jeffreys' (1939) claim that in the absence of a reason to consider one hypothesis more likely than another the probabilities are equal—to the situation in which the question is what probability to assign to the drawing of a ball of a specified color from an urn containing balls of three colors, say blue, red, and yellow, in unknown proportions:

> Let us consider as a starting hypothesis that the first ball we draw from the urn will be blue. According to Jeffreys' (and Laplace's) statement of the principle of indifference, if the question is whether the first ball will be blue or not blue, we must assign equal probabilities to both these hypotheses; that is, each probability is 1/2. If the first ball is not blue, it may be either red or yellow, and again, in the absence of knowledge about the actual proportions in the urn, these two have equal probabilities, so that the probability of each is 1/4. But if we were to start with the hypothesis that the first ball drawn would be, say, red, we would get a probability of 1/2 for red. Thus Jeffreys' system as it stands is inconsistent. (p. 132)

Carnap suggests that "one of the fundamental questions to be decided is whether to accept a principle of indifference, and if so, in what form. It should be strong enough to allow the derivation of the desired theorems, but at the same time sufficiently restricted to avoid the contradictions resulting from the classical form" (p. 134).

The principle of indifference is problematic even—perhaps especially—when applied to situations that do not have the complication of alternative plausible interpretations like that described by Carnap. Consider the workhorses of probability theory: coins, dice, and cards. We assume that the (fair) coin is equally likely to come up head or tail, that the (fair) die is equally likely to show any of its six faces, and that each of the possible hands of, say, 13 cards (as in bridge) is equally likely to be dealt from a well-shuffled deck. But why should we make this indifference assumption in any of these cases? Ayer (1965) raises this question and proposes that, in the absence of experience, we have no reason to do so: "Antecedently to experience ... we have no more reason to expect that the results of tossing coins or throwing dice will conform to the a priori probabilities than that they will deviate from them. The reason that we think that results that are highly improbable in this sense call for a special explanation is that they are empirically abnormal" (p. 51). As a matter of historical fact, statistical regularities were observed in nature before any rational basis for predicting them had been developed; it was the observation of them that prompted the effort to provide a theoretical explanation. It is not clear that any explanation has been found; it appears, as Ayer seems to be claiming, that we expect statistical regularities because that is what we have observed and we can say no more than that.

Probability and Uncertain Knowledge

> Probability in its most general use is a measure of our degree of confidence that a thing will happen. (Tippett, 1941/1956, p. 1485)

Whatever else it may be, the idea of probability is a reflection of the fact that our knowledge of the world is often uncertain and incomplete. Because our individual knowledge bases about specific aspects of the world differ, the probabilities we assign to possible events should sometimes differ also. If a fair coin were about to be tossed, you and I would undoubtedly agree that the probability that the toss would result with a head is .5. If the coin had already been tossed, but neither of us had seen the outcome, we both would again consider the probability of head to be .5. If you now looked at the coin but I did not, the probability of head for you would immediately become either 0 or 1, but for me it would remain at .5. Your uncertainty about this aspect of the world has been resolved by the information you received from your observation, but my uncertainty remains. Or if one of us knew the coin tosser to be using a trick coin weighted so as to produce a head on about 75% of the tosses, and the other did not have this knowledge, again the probability of the outcome of the upcoming toss would be different in our minds.

In this book, a number representing a probability will be taken to reflect a belief regarding the likelihood of the occurrence of a particular event. Generally it will be assumed that such beliefs are based on, and are consistent with, either assumptions of physical symmetry (as would be involved in coin tossing or dice throwing) or objective relative frequency data (as, for example, in actuarial statistics) when such data are available. When notions of symmetry are not relevant or relative frequency data are not available (as when estimating the probability of a third world war before the middle of the 21st century), such a number will be taken to reflect the judgment of an individual or group, the basis of which may or may not be possible to make explicit.

For purposes of describing experimental investigations of reasoning under uncertainty, it may not be necessary to make much of the distinction between relative frequency and probability. To be sure, some investigators have had people estimate relative frequencies and some have had them estimate probabilities, but typically the basis for judging the accuracy of the probability estimates (when that has been a concern) has been relative-frequency data. When, for example, people have been asked to estimate the probability that one will be the victim of some type of accident, it has been assumed that the true probabilities are seen in past-incidence statistics. In this book, when discussing experimental work on reasoning under uncertainty, I will follow the convention of using the same terms (relative frequency, probability) as have the investigators whose work is cited, and generally will not make anything of the differences in terminology.

Until the development of the theory of quantum mechanics early in the 20th century, the need to make use of statistical probability was assumed to be a reflection of the limitations of our powers of observation and understanding of physics. The outcome of the toss of a coin, for example, was assumed to be predictable, in principle, though not in practice, given the current state of understanding of the interacting causal forces. With the development of the theory of quantum mechanics came acceptance of the idea that some properties of the universe are inherently probabilistic, and are not predictable, except statistically, even in principle. In quantum theory, distributions of quantum events are predictable, but individual events are not. The intensity of a field at a particular point is said to equal the probability of the occurrence of the field's associated quantum particle at that point; whether it will actually be observed at that point can be predicted only probabilistically until the observation is made.

The debate between those who hold a frequentist view of probability and those who hold a subjectivist view is unlikely to be resolved soon; there are knowledgeable people on both sides of the divide who see their position as obviously correct and who fail to understand how anyone could find the opposing view tenable. Fortunately, the debate represents little threat to probability calculus as a mathematical discipline; the mathematical theory of probability, especially as axiomatized by Kolmogorov, appears to be immune to disagreements about what probability really is. It is remarkable that the theory has proved to be so practically useful in so many contexts, despite the continuing lack of consensus regarding what the basic concept means, but a similar observation can be made about mathematics more generally.

G. Shafer (1993) argues that an axiomatic theory of probability can be based on axioms pertaining to any of three ideas—knowledge of the long run, fair price (or fair odds), and warranted belief—that the concept of probability ties together in what he calls a circle of reasoning. But whichever idea is selected as foundational, one cannot do full justice to probability without involving the other two ideas as well; the three were inextricably intertwined in the emergence of mathematical probability. Historically, Shafer attributes the development of the theory of fair price in games of chance to Pascal, Fermat, and Huygens in the 1650s, and the step from fair price to probability during the following 50 years primarily to Jacob Bernoulli, who also proved the law of large numbers, which is fundamental to knowledge of the long run. Knowledge of the long run was brought back to fair price by Condorcet in the 1780s, thus making the circle complete. The contemporary thinker can enter this circle at any point, Shafer suggests, but getting the ideal picture of probability requires that one appreciate all three ideas and the connections among them.

WHAT IS AN IMPROBABLE EVENT?

Some of the concepts relating to probability theory are somewhat difficult. Arguably, many, if not most, of them are relatively simple. Confusions sometimes arise unnecessarily, however, because of imprecision in the use of language.

For this reason, it often is hard to tell whether certain confusions are simply the results of imprecise use of language or have deeper conceptual roots. Consider, for example, the basic idea of an improbable event.

Specific Events Versus Events With Specific Properties

Imagine a bag containing 10 marbles, each with a unique number on it between 1 and 10 inclusive, and suppose someone were to draw them from the bag, while blindfolded, one at a time (sampling without replacement). A drawing in the order 8, 2, 3, 6, 10, 4, 1, 5, 9, 7 would not be considered remarkable, whereas to most people a drawing in the order 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 undoubtedly would. Given that each of these specific orderings is equally improbable, why should one of them be readily accepted as the result of a random draw whereas the other one is likely to raise suspicions about the drawing process?

Or consider two sequences of coin tosses with 10 tosses in each one. Suppose one sequence is H H T H T T T H T T and the other is 10 consecutive heads. Again, most people would not find the first sequence remarkable, but the second one would lead them to wonder whether the coin had a head on both sides. In fact, assuming a fair coin, the two sequences are equally likely, or—to be more precise—equally unlikely, inasmuch as the probability in each case is 1 in 2^{10}, or 1 in 1,024.

These examples illustrate the importance of distinguishing between a specific event and an event with specific properties (Nickerson, 2002). Consider again the coin-tossing case. The first sequence *looks* more random, and consequently seems more likely, than the second because one expects the sequence to have heads and tails in roughly equal proportions. We run into difficulty here when we fail to distinguish the event H H T H T T T H T T from the *set* of events having the property, say, "from four to six heads." Inasmuch as there are 672 ways of obtaining from four to six heads in 10 tosses, getting *some member* of this set is 672 times as likely as getting the single member of the set "10 consecutive heads," but getting *the particular member* that was obtained is not. Similarly, returning to our first example, the drawing of 8, 2, 3, 6, 10, 4, 1, 5, 9, 7 is no more likely than the drawing of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, but the probability of drawing some sequence that is not perfectly ordered, of which there are many, is very much greater than the probability of drawing the single one that is.

It is easy to imagine one saying, at this point, "I see the distinction, but I would still be more surprised to see 10 tosses of a coin result in 10 heads than to see them result in H H T H T T T H T T." Would such an attitude be justified? The answer is, it depends on what, precisely, one means. If one means that one would be less surprised to see a result containing four heads and six tails than to see one with all heads, it is justified. If one means that one considers the *specific* result H H T H T T T H T T to be more likely than 10 consecutive heads, it is not.

Another way to make the distinction is to contrast irregularity with randomness (or regularity with nonrandomness). The fact that a sequence is irregular is not compelling evidence that it was produced by a random process, nor is the fact that it is regular proof that it was not produced by a random one. It will perhaps be readily conceded that an irregular sequence can be produced by a nonrandom process; what may be less apparent is that a regular sequence can be produced by a random one. The tossing of a fair coin *can* yield 10 heads in a row; it is unlikely that it will do so, but not impossible. Such an outcome might be described as "chance regularity." It is much more likely that the tossing of a fair coin will result in a mix of heads and tails. So it is reasonable to associate irregular sequences with randomness and regular sequences with nonrandom processes, as a general rule, even though any *specific* irregular sequence is no more likely than a comparable specific regular one.

Twice within a period of a few months a few years ago, the Massachusetts and New Hampshire lotteries turned up the same (four-digit) number on the same day. The first time this happened, the event was considered sufficiently improbable to merit a prominent article in the *Boston Globe*. The article misreported the chances of the two lotteries producing the same number on the same day as being 1 in 100 million. In fact, they are 1 in 10,000. What the article had reported was the probability that the lotteries would both turn up a *specific* number (say, 2734). The chance of that happening is indeed 1 in 100 million, but because there are 10,000 numbers (between 0000 and 9999 inclusive) that the two lotteries could both pick, the chance of them picking the *same* number (irrespective of what the number is) is 10,000 times 1 in 100 million, or 1 in 10,000.

One can say of any specific pair that the chance of that particular combination coming up is 1 in 100 million. This being so, why is it that the one outcome, say 2734 and 2734, attracts attention and another, say 6135 and 2864, does not? The answer is in the way in which outcome of interest is defined. In the aforementioned example, the outcome of interest is the occurrence of any two numbers that are identical, which is a rare outcome (1 chance in 10,000) relative to the occurrence of any two numbers that are not identical (9,999 chances in 10,000). Both of these outcomes are sets composed of disjunctions of many possible combinations. The outcome *identical numbers* includes the events 0000–0000, 0001–0001 ... 9999–9999. The outcome *nonidentical numbers* includes all other combinations of which there are 9,999 times as many.

I have belabored the distinction between specific events and events with specific properties because I believe it to be not only important to an understanding of many probabilistic phenomena but also extremely easy to forget, or to obscure with the careless use of language. So we may refer to an event as being more or less probable when what we should say is that observing *some* event with certain properties of interest in common with the one in question is more or less probable. This type of confusion may help account for a variety of results that have been obtained in experimental studies of statistical or probabilistic reasoning.

If one understands this distinction, one should recognize that any sequence of numbers in a fair lottery is as likely to occur as any other. Thus in a lottery in which five numbers are drawn, without replacement, from, say, the numbers 1 through 20, the sequence 1, 2, 3, 4, 5 is precisely as likely to occur as is the sequence 11, 5, 20, 3, 16. This being true, is it rational to have a preference for the second sequence over the first? An argument can be made either way. On the one hand, one might prefer to have a ticket with the second sequence rather than one with the first, on the grounds that suspicions that would be likely to arise if the former sequence were drawn would not be evoked by a drawing of the latter one, even though each is as likely (or unlikely) as the other. On the other hand, if one assumes that people would be disinclined to select a regular sequence like 1, 2, 3, 4, 5, one might prefer to select it with the idea that if one won, one would be unlikely to have to share the prize with other winners.

The distinction between specific events and events with specific properties is relevant to an understanding of the second law of thermodynamics, according to which the entropy (randomness) of a closed system tends always to increase, which is to say that the natural tendency of any closed system is to proceed from more orderly to less orderly states, or in probability terms, from a less probable state to a more probable state. The principle is often illustrated with the example of the distribution of gas molecules in a closed container. Suppose we have a container with molecules of, say, neon and argon, in equal numbers and begin with all the neon molecules congregated on one side of the container and the argon ones on the other. We would say that this is a highly ordered (improbable) state of affairs. If we look at the container again after it has been left alone for a while, we will find that the neon and argon molecules are considerably more mixed—the longer we wait the more thoroughly they will be mixed—and we would say that the situation has less order (a more probable distribution) than it had before. If, on the other hand, we begin with the molecules mixed, and inspect the container after leaving it alone for a while, we are very unlikely to discover that the neon has migrated to one side of the container and the argon to the other—that the distribution has changed from a less orderly (more probable) state to a more orderly (less probable) one.

But suppose that every molecule were identifiable—we had neon molecules Ne_1, Ne_2, Ne_3, and so on, and the argon molecules were similarly tagged—and we wished to describe the distribution of the *individual* molecules in the container. Inasmuch as this description will require the specification of the *x, y, z* coordinates of every molecule, the description will be equally complex whether the neon and argon are segregated or thoroughly mixed. Moreover, if we assume that all possible distributions are equally likely, say after the container has been undisturbed by any outside influence for some time, any specific mixed distribution has precisely the same probability of being obtained as any specific segregated distribution. The probability that some mixed distribution will be obtained is much greater than the probability that some segregated distribution will be, simply because there are many more possibilities of the former type. The situation is completely analogous in this respect to the coin-tossing and number-drawing events considered earlier. Any specified sequence of tosses or drawings is as likely as any other, but the probability of getting a sequence of tosses that contains a mixture of heads and tails is greater than that of getting one that contains only heads, and the probability of drawing numbers in a mixed order is greater than that of drawing them in their natural order, simply because there are more possibilities of the latter type than of the former in both cases.

What does this all mean with respect to our understanding of the second law of thermodynamics? It means that whether we see entropy increasing depends on the level of detail at which we choose to describe the system. Physicists refer to the coarseness of the grain of a description; the coarser the grain, the less detailed—hence less complex—a description is said to be: "Entropy is a useful concept only when a coarse graining is applied to nature, so that certain kinds of information about the closed system are regarded as important and the rest of the information is treated as unimportant and ignored" (Gell-Mann, 1994, p. 371). In the case of the example of the distribution of gas molecules, entropy would be seen to increase, given a description of the situation that is relatively coarse-grained, but not given a description that is sufficiently fine-grained to track the movements of individual molecules. In terms of the distinction that has been made here between specific events and events with specific properties, we might say that entropy would be seen to increase if one's description focuses on events with specific properties—such as the clustering or dispersion of molecules of specific types—but not if it focuses on the specific distribution of individually identifiable molecules.

The results of several experiments suggest that the distinction between specific events and events with specific properties is not clearly understood by many people. In one study, high school students were given the following problem: "All families of 6 children in a city were surveyed. In 72 families the exact order of birth of boys and girls was G B G B B G. What is your estimate of the number of families surveyed in which the exact order of births was B G B B B B?" (Kahneman & Tversky, 1972, p. 432). About 80% of the subjects in this study judged the latter sequence to be less likely than the former; the median of their estimates for the number of families with this birth order was 30. Other investigators who have obtained similar results include Tune (1964) and Wagenaar (1970).

Assuming the probabilities of male and female births to be equal, the two birth sequences G B G B B G and B G B B B B are equally probable (and the same also as the orders BBBBB and GGGGG). Obtaining a sequence composed of about half boys and half girls, however, is considerably more likely than obtaining one composed almost entirely of boys. If people are confused with respect to this distinction, they may, in some cases at least, judge the probability of obtaining *sequences with specific properties* similar to those of the given sequences, rather than the probability of obtaining those *specific sequences.*

Kahneman and Tversky (1982b) note that sometimes the occurrence of the more probable of two outcomes of a random process can be seen as more surprising than the occurrence of the less probable outcome. They give the example of a fair coin being tossed 40 times: The most likely result, in terms of numbers of heads and tails, is 20 heads and 20 tails, but some people are more surprised by this outcome than by the result 22 heads and 18 tails. Kahneman and Tversky give two possible explanations of this reaction. First, it may be that the 22-18 split is considered more representative of a random sequence than the 20-20 split. Second, perhaps one's expectation for the outcome of the coin tossing is for an *approximately* even split and, if so, the outcome 20-20, which is an exactly even split, would be perceived as a low-probability event.

This situation, like the one involving birth order, is complicated by the imprecision and ambiguities of language. To determine whether the surprise registered by the people in this example is warranted, one must have a clear understanding of what constitutes the outcomes of interest in their minds. The outcome 20 heads and 20 tails in 40 tosses of a coin is more likely ($p = .125$) than the outcome 22 heads and 18 tails ($p = .103$) and if these are the two outcomes that are being contrasted, then one should not be more surprised to see the first one occur than to see the second one do so. Outcomes of interest can be defined in other ways in this situation, however. The probability of a 22-18 split (i.e., 22 heads and 18 tails *or* 22 tails and 18 heads) ($p = .206$) is greater than the probability of a 20-20 split. And the outcome "approximately even split" (say from 23-17 to 17-23, excluding 20-20) is very much more likely ($p = .606$) than a precisely even split. If the people in Kahneman and Tversky's example had in mind one of these types of comparisons, then their surprise at the occurrence of a 20-20 outcome was less clearly unjustified.

In general, failure to make a distinction between specified events and events with specified properties is bound to lead to confusion in the assessment of probabilities. And this is a distinction that we should not assume most people make spontaneously. Before judging a reaction to the outcome of a probabilistic process to be irrational or otherwise inappropriate, however, we need to know what is the nature of the *event,* in the mind of the observer, to which the reaction is made. A reaction that could be viewed as irrational if the event of interest is conceived in one way might be quite reasonable if conceived in a different way.

What Establishes the Probability of an Event?

Fair coins, if tossed repeatedly, will come up heads approximately half of the time. No one, no matter what one's understanding of what probability "really is," doubts that this is so. But why is it so? What determines that heads and tails will occur with about equal frequency? This question pertains to physical or aleatory probability, and it is one that, in one form or another, has perplexed probability theorists, and other people who have thought about such things, for a long time.

Consider again Bernoulli's limit theorem. What gives us confidence that what this theorem claims—that "in the limit" the relative frequencies with which specific events are observed in a sample will be arbitrarily close to the probabilities of those events—is true? We cannot check it empirically, because we cannot examine a sample that is "close to infinity" in size. Moreover, suppose we could do so. If we found that the observed relative frequencies of specific events in such a sample differed from the assumed probabilities of those events, we would revise our assumptions regarding the probabilities. If we found, for example, that a large number of tosses, say a million, of a coin yielded heads and tails in the ratio 55:45, we would conclude not that the limit theorem is wrong but that the probability of that particular coin coming up head is not .5 but closer to .55. And we would do so on the strength of Bernoulli's limit theorem or reasoning of a similar kind. But is this not circular?

What are we to say to Gigerenzer et al.'s (1989) claim that Bernoulli's limit theorem "created a model of causation that was essentially devoid of causes" (p. 29)? Is it true, as these writers suggest, that this new model "abandoned all search for mechanisms, for the hidden springs and principles that ran the clockwork of the world" (p. 30)? "In Bernoulli's urn model," they argue, "numbers generated numbers; the physical processes by which they did so were wholly inscrutable" (p. 30). Are we closer to an understanding of what probability really is today than when Bernoulli's theorem and urn model were originally articulated?

Early literature on probability contains references to possibilities, propensities, proclivities, and facilities, but such terms have little explanatory power. Sometimes the circularity of attempts to define or account for probability was glaring, as when Laplace, in his first paper on the topic, noted that probability could be defined in terms of a ratio among cases, so long as the cases are equally probable (*Oeuvres*, VIII, p. 10; noted in Hacking, 1975, p. 131). Fortunately, as already noted, it is not necessary to be able to explain probability in order to be able to use the concept to good effect.

The Ubiquity of the Improbable

> It would be very unlikely for unlikely events not to occur. (Paulos, 1990, p. 37)

> The essence of chance is that anything is possible, even the improbable. (Ekeland, 1993, p. 113)

> Just how great do the odds against a particular outcome have to be before one can be confident that the outcome is not the result of chance? (Devlin, 2000b, p. 14)

Poincaré (1913/1956) described the birth of a great man as the greatest bit of chance:

> It is only by chance that meeting of two germinal cells of different sex, containing precisely, each on its side, the mysterious elements whose mutual reaction must produce the genius. One will agree that these elements must be rare and that their meeting is still more rare. How slight a thing it would have required to deflect from its route the carrying spermatozoan. It would have sufficed to deflect that a tenth of a millimeter and Napoleon would not have been born and the destiny of our continent would have been changed. No example can better make us understand the veritable characteristics of chance. (p. 1392)

Poincaré's (1913/1956) point is interesting, but it understates the improbability of the existence of Napoleon, or of any particular person by a good bit. The existence of each one of us is so incredibly improbable that one can become dizzy thinking about it. Consider. Every human cell has 23 pairs of chromosomes. When a reproductive cell divides during meiosis in preparation for possible union with another reproductive cell, each of the resulting haploid cells has 23 chromosomes, one from each of the 23 pairs. Assuming that each member of any given pair is as likely as the other to end up in a particular haploid cell, there are 2^{23}, or more than 8 million different haploids that can be generated by the meiosis of one diploid cell.

When a sperm cell and an egg cell are united to form a new diploid, the sperm is 1 in 2^{23} possibilities, as is the egg; therefore, the combination of chromosomes in the new diploid is 1 of 2^{46} or more than 70 trillion possibilities. Thus given the existence of individuals with the genetic endowments of your parents, the probability that their union would result in an individual with precisely your genetic makeup is 1 in about 70 trillion. This number is more than

10,000 times as large as the current world population and probably several thousand times as large as the number of human beings who have ever lived. Which is to say that the reproductive cells of two human parents are potentially capable of generating the genetic blueprints of more different individuals than have walked the face of this planet, without ever making a duplicate.

But of course these figures do not begin to reflect the uniqueness of a human being, because they assume (a) the existence of an individual with the person's mother's genetic endowment, (b) the existence of an individual with his or her father's genetic endowment, and (c) the meeting and mating of those two particular individuals. Given the existence somewhere among more than 6 billion people 2 with specific genetic makeups, the probability of their meeting and mating must be vanishingly small. And an attempt to estimate the a priori probabilities that such people should come to exist involves us in a very lenghty regress.

Another way to look at uniqueness is to consider the makeup of an individual at the level of the gene. Heterozygosity is defined as the average proportion of gene loci in the DNA molecules of an individual that are occupied by two alleles. In man the average heterozygosity is estimated to be about 6.7%. Assuming an individual has about 100,000 gene loci, this means that about 6,700 of those loci would be occupied by different alleles and that consequently an individual has the potential to produce $2^{6,700}$, or $10^{2,017}$, unique germ cells (Ayala, 1978). This is an unimaginably large number. If one were to subtract from this number the number of all human beings who have ever lived, the remainder would be (in round numbers) $10^{2,017}$. So much is clear: The probability that an individual precisely like you should ever have come into existence is so close to zero to be, for all practical purposes, indistinguishable from it. But then, as Thomas (1979) points out: "Uniqueness is so commonplace a property of living things that there is really nothing unique about it" (p. 2).

Uniqueness, or unlikeliness, is a common property not only of living things but of the physical world as well:

> Rarity by itself shouldn't necessarily be evidence of anything. When one is dealt a bridge hand of thirteen cards, the probability of being dealt that particular hand is less than one in 600 billion. Still, it would be absurd for someone to be dealt a hand, examine it carefully, calculate that the probability of getting it is less than one in 600 billion, and then conclude that he must not have been dealt that very hand because it is so very improbable. (Paulos, 1990, p. 54)

The fact that any particular hand is so improbable is undoubtedly one of the reasons people can play such a game as bridge for many years without becoming bored with it. One can easily play the game regularly for many years without ever seeing the same hand twice.

Paulos' claim that it would be silly to decide that one could not have been dealt the bridge hand that one holds on the grounds of its improbability is unlikely to provoke much debate. The bridge hand illustration makes the case too for the observation by Devlin (2000b) that "simply computing the probability of an event is not enough to decide whether some phenomenon is the outcome of chance or design" (p. 14). But suppose now that someone predicts the cards that her *next* hand will contain and it turns out that she is dealt precisely the predicted cards. The probability that the prediction would prove to be accurate, assuming a random deal, is less than 1 in 600 billion, the same as the probability of getting any particular unpredicted hand, but in this case we would all be surprised indeed that the prediction came true and we would surely believe that the deal was not quite as random as it was supposed to be.

Why are we surprised in the second case and not in the first? *Should* we be surprised in the second case and not in the first? What evokes the surprise, or what should do so, is not the fact that a low-probability hand has been dealt but that in the second example one was able to predict the dealing of a particular low-probability hand. If we registered surprise at every occurrence of a low-probability event, we would exist in a state of constant amazement, because all events are, from one or another perspective, low-probability events. The ability to predict the occurrence of *specific* low-probability events, however, is something we do not, generally speaking, have.

To make the same point in slightly different terms, we can note that life is full of situations in which we can say that an improbable event is certain to occur—that the probability that an improbable event will occur is 1. Such events are represented by a lottery with, say, 1,000,000 participants. Assuming the lottery is fair, the probability that any specific participant will win is very small, 1/1,000,000, close enough to 0 to be considered 0 for practical purposes. The probability that *someone* will win, however, is 1. Some individual will feel very fortunate after the fact and may be inclined to wonder "why me?" But it had to be somebody; the only way to avoid that is to not have the lottery, and most of the real-life events for which this is symbolic do not give us that option.

I have argued that the chance occurrence of a low-probability event should not necessarily evoke surprise, but that the chance occurrence of a *predicted* low-probability event should; the occurrence of a predicted low-probability event is likely to make us wonder whether it really occurred by chance. This is not the whole story, however; sometimes we are truly surprised by low-probability events even if they were not predicted, and it is hard to argue convincingly that we should not be. Consider again the bridge hand. We are not surprised when we are dealt a hand that has less than 1 chance in 600 billion of being dealt, but we would be surprised if dealt a perfect hand—13 cards of the same suit—even though the probability of receiving *that particular* hand is just the same as that of getting any other particular hand.

Why are we more surprised at getting a perfect hand than at getting any particular hand? Again we come back to the distinction between specific events and events with specific properties. Our surprise at being dealt a particular bridge hand is determined, at least in part, I suspect, by the relative size of the set of possible hands that share salient features with the hand in hand. There are only four members of the set of perfect hands, whereas most specific hands have features in common with many other hands. The probability of getting a hand with, say, four cards in each of two suits, three in a third, and two in the fourth is very large by comparison with the probability of getting one with all cards in the same suit. The probability of getting a hand with a mix of face and number cards is much larger than the probability of getting one with all face cards. And so on. We need not suppose that one can calculate, or even accurately estimate, the probabilities involved in order to have some rough feel for them. People who play the game frequently are likely to acquire from experience an approximate knowledge of the relative frequency with which hands with certain properties of interest occur, and even the most inexperienced novice is likely to recognize that a hand with a mix of suits is much more probable than one with all cards of the same suit.

The distinction between small-probability events that are noticed after the fact and those that are predicted in advance is critically important for many purposes; however, I do not mean to suggest that low-probability events that are noticed after the fact should always be dismissed as chance events. If I toss a coin 10 times and it comes up head every time, I am surprised, whereas if it comes up TTHTHHHTHT, I am not, despite the fact that I recognize the probability of each sequence to be precisely the same—about .001. The all-heads sequence makes me wonder about the fairness of the coin and the other one does not. Are my different reactions justified? The reader who feels they are not might imagine tossing the coin 100 times, or 1,000 times, and asking whether a sequence of all heads would be more surprising than *a particular* sequence with a mix of heads and tails, despite the fact that any two specific 100-toss (or 1,000-toss) sequences are equally probable.

The argument I want to make is that discovered low-probability regularities can be legitimate bases for surprise and that they can serve as reasonable stimuli for hypothesis formation and can motivate searches for causal explanations. Getting all heads in even only 10 tosses of a coin justifies, in my view, some suspicion that something peculiar is going on. I would want to entertain—and test—some hypothesis other than that the coin is fair. The observation of the 10 heads does not prove that the tossing was a nonrandom process, but it raises the suspicion and rightfully prompts the search for a causal explanation. When does perceived low-probability structure warrant investigation? When should it be taken as suggestive evidence of a nonchance effect? When does it beg a causal explanation?

These are difficult questions, and challenges to the statistical decision making, a subject to which we return in chapter 10. It suffices to note here that apparent structure *can* happen by chance and apparent structure *will* happen by chance. Search for structure among outputs of a random process is bound to succeed. It does not follow, however, that all apparent structure should be dismissed as effects of chance.

How Does One Prove a Probability to Be Wrong?

Suppose I were to say that the probability of getting 10 heads in a sequence of 10 coin tosses is approximately .001 and that you then set out to check the accuracy of this claim by doing a set of trials, each consisting of 10 tosses, counting the number of heads obtained in each case. Imagine now that you get 10 heads in a row on your very first trial. Does this disprove my claim that the probability of getting 10 heads in a row is about .001? The claim was not that it was *impossible* to get 10 heads in a row on any particular trial, but only that such an outcome was improbable. Suppose you got 10 heads in a row on each of your first three trials. Surely, this would prove conclusively that my claim was wrong.

But again, the claim does not make that event impossible either; assuming the claim is true, three consecutive trials of 10 heads is to be expected about one time in a billion, on the average, but maybe this it that one time—it is not impossible. We can keep on in this vein; no matter how many times you toss 10 heads in a row, I can say that my claim that a single such outcome is about .001 does not rule out the possibility of a series of such outcomes of any given length. To be sure the probability becomes vanishingly small, but no smaller than that of any particular outcome. I could always argue, could I not, that the event, improbable though it is, is consistent with my original claim. You would long since have come to the conclusion that there was something wrong with the coin, and so would I, but it is still the case that the possibility of such a surprising outcome cannot be said to be inconsistent with my claim.

For practical purposes, we act, in some instances, as though the occurrence of low-probability events is indeed impossible, or at least we take such occurrences as evidence that our assumptions about their probability of occurrence were wrong. The explosion of the Challenger on the 25th launch of a space shuttle, for example, has been taken by some as evidence that the National Aeronautics and Space Administration's (NASA, 1985) estimate that the probability of such an accident was about 1/100,000 was grossly in error. In fact, if the probability of such an accident really was 1/100,000, this would not rule out the possibility of one occurring on the 25th launch—or on the 1st launch, for that matter—but this argument is not likely to count for much—and perhaps should not—in the practical world of assessing risks to life. It is possible that NASA's estimate was correct and that the Challenger disaster had a 1-in-100,000 chance of occurring as it did; alternatively, the probability estimate may have been overly optimistic. The choice is not incidental, inasmuch as how much effort is put into reducing the probability of a future accident is likely to depend on which of these views is taken.

In using statistics to test scientific hypotheses, the convention to reject a null hypothesis—to consider it to be false—if the chance probability of obtaining the observed outcome is considered to be less than a specified small value, say .01. Strictly speaking, the logic of statistical hypothesis testing does not allow one, when one obtains a "statistically significant" result, to rule out the *possibility* that the result was obtained by chance; but the convention is to behave as though it did so.

SO WHAT REALLY IS CHANCE?

> Paradoxically, chance lies at the root of most of the uniformities of the world we are familiar with. (Barrow, 1990, p. 297)

> The existence of chance, far from opposing the order of the universe, manifests ever more cogently the existence of order. (Nogar, 1966, p. 292)

Chance has been evoked many times in the preceding discussion; it is time to take stock of this concept. We speak of "chance devices," "chance events," and "chance variations." We refer to "effects of chance," noting that this or that event was "due to chance" or "happened by chance." We refer to many games as "games of chance." The importance of the concept to mathematics and science is seen in references to the "mathematics of chance," the "theory of chance," the "doctrine of chances," "the science of chance phenomena," and the, some would say incongruous, term "the laws of chance."

Chance is a controversial concept. As I see it, the word has several connotations as used in the scientific and mathematical literature. I note three:

- *Chance as a cover for undetermined physical causes.* When I say that whether a coin I am about to toss will land head or tail is a matter of chance, what do I mean? I do not mean that the outcome of the toss is not determined by the laws of physics. I mean only that I am unable to determine what those laws will dictate in this particular instance, because I do not know the parameters of the situation essential to make that determination. I do not know, for example, how high the coin will be thrust into the air, the rate at which it will revolve, the elasticity of the surface on which it will land, and so forth. As to why the two possible outcomes are likely to occur with roughly equal relative frequency in a large number of tosses, I assume that the values of the determining parameters will change haphazardly from toss to toss and that there is no reason why they should consistently favor one outcome over the other, but I consider the outcome to be determined by the physics of the situation in all cases. If allowed to toss the coin only a little way (say a few inches) into the air and to let it land on a soft surface from which it would be unlikely to bounce, I might, with lots of practice learn how to control the toss well enough to make the coin land one way considerably more often than the other.

- *Chance as a cover for ignorance.* Suppose a friend has tossed a coin and knows the outcome, but has not told me what it is. This situation is very different from the preceding one. To my friend the probability that the coin came up head is either 0 or 1, but from my perspective, I consider the probability that it was head to be .5. One might say that to me the outcome of the toss is still a matter of chance, and this would be entirely in keeping, I think, with the way the word is often used.
- *Chance as an explanation.* Sometimes the intention in evoking chance appears to be to explain some phenomenon of interest. When one says this or that event was "due to" chance, happened "by" or "because of" chance, or "was governed by" chance, one may mean simply that the event was caused by factors that cannot be identified or that are beyond one's control. This is essentially the first connotation mentioned earlier. However, another interpretation that may be given to such a statement is that chance was the cause of the event. This use of the concept is not appropriate or helpful, as I see it.

In my view, chance is a descriptive concept and never an explanation. We may say that under specified conditions, certain events are best described as chance events—that the various possibilities have equal probability of occurring—and this observation may be very useful, but simply calling it chance does not explain why the behavior is such as it is. The atoms of a radioactive substance are constantly decaying, the rate of decay differing over a very large range from substance to substance. We say that precisely which of the atoms will decay during any particular period—or whether any specific atom will do so—is a matter of chance. But by calling it a matter of chance adds nothing to the observation that every remaining atom has the same probability of decaying immediately, and that this does not change over time. And the invocation of chance does not explain why the decay process is as it is.

Chance is one of those terms that lend themselves to the fallacy of reification—the uncritical assumption that a name must name some *thing*. In invoking the name, we convince ourselves that we have explained some phenomenon of interest. But in what sense have we done so? Attributing the behavior of something or someone to chance is analogous to saying that a person picks fights because of a quarrelsome nature. In this respect, the concept of chance is on a par with that of gravity. Saying that bodies attract each other because of the force of gravity really says no more than that bodies attract each other. The inverse square law of gravitational attraction appears to be a universal law, something that characterizes the behavior of matter everywhere and everywhen, and the observation that this is the case is immensely useful, but stating the law does not constitute an explanation of the why of it. Just so with the concept of chance. Many phenomena, especially involving the behavior of aggregates, are well described as chance phenomena, or as phenomena that obey the laws of chance; such descriptions can be immensely useful, but they leave unanswered the question of why such lawfulness is observed.

SUMMARY

Some conception of chance dates to antiquity, as evidenced by such practices as the drawing of lots, the rolling of bones, and other forms of gambling. Despite this fact, the development of a quantitative theory of probability did not begin in earnest until the 17th century. Accounts of the thinking of the early developers of this theory reveal struggles between conflicting intuitions regarding what the answers to specific questions of probability should be. Such conflicting intuitions were often resolved by consensus, but not always.

As to what probability really means, differences of opinion persist. Frequentist (or statistical) and subjectivist (or judgmental) connotations represent one, and perhaps the most important, distinction in both historical and contemporary views. The development of probability theory as a mathematical discipline seems not to have been impeded by the lack of consensual closure about meaning; the theory has been, and continues to be, used to good effect in numerous areas of practical application.

The concept of chance, which is central to probability theory, also has several connotations in the literature, and what, precisely, it should be taken to mean is still a matter of debate. That what are generally considered to be chance events are lawful, which is to say predictable, in the aggregate is beyond dispute, thus the peculiar notion of "laws of chance." But chance is not an explanatory concept, at least in the view of the writer; why chance events are lawful remains an unanswered question.

CHAPTER 2

Randomness

Probability is the branch of mathematics that describes randomness.

—Moore (1990, p. 98)

The puzzle is this, if randomness is a product of ignorance, it assumes a subjective nature. How can something subjective lead to laws of chance that legislate the activities of material objects like roulette wheels and dice with such dependability?

—Davies (1988, p. 31)

Whatever your views and beliefs on randomness—and they are more likely than not untenable—no great harm will come to you.

—Kac (1983, p. 406)

The concept of randomness is fundamental in probability theory. At one level, the concept is relatively simple and intuitively easy to grasp. The selection of a number between 1 and 10, inclusive, would be considered a random selection if it were such that every number from 1 to 10 had an equal chance of being selected. However, specification of a selection procedure that guarantees this equal-chance requirement is considerably more difficult than describing the requirement. More generally, the concept has proved to be more than a little elusive, and it remains enigmatic, despite the fact that it has proved to be very useful in many contexts.

WHAT IS RANDOMNESS?

Randomness must be easy to define—P. J. Davis and Hersh (1981) point out that there are a good dozen different definitions of a random sequence—but it is not easy to find a definition with which experts agree. There are, however, certain concepts that one encounters often in discussions of randomness and that characterize properties that a random set or sequence is expected, at least by some observers, to have. Among the more common of these properties are equal representation, irregularity or unpredictability, and incompressibility.

Some Conceptions

During the 1870s, William Shanks published the value of π to 707 places, a prodigious feat, given that the computation was done entirely by hand. Three quarters of a century passed before someone produced, with the help of computing machinery, an approximation with a larger number of digits. Before this time, the last 200 or so digits of Shanks' approximation had been suspect, because some digits were represented noticeably more than others. Inasmuch as π is a transcendental number, the digits should be randomly distributed, and therefore should appear with about equal frequency in any sizeable part of its decimal approximation. As it turned out, Shanks' approximation was indeed incorrect after about the 500th decimal place (Ogilvy & J. T. Anderson, 1966).

This story illustrates the equal-representation conception of randomness: Given a process that selects repeatedly (with replacement) from a finite set in such a way that every member of the set has the same chance as every other of being selected on each occasion, we would expect every member of the set to be represented approximately the same percentage of times among the selected items after a large number of selections.

This property is easily misunderstood. If an urn contains three black balls and one white one, we do not expect black and white balls to be represented equally in a large number of random selections, with replacement, from the urn. What we mean by equal representation is that in a large number of random selections, each *ball* will be represented approximately the same number of times, or that on any given draw every ball, independently of color, has the same chance of being drawn.

A random set or sequence is said to be irregular in the sense that it is relatively unstructured. I say "relatively unstructured" because structure or regularity is a matter of degree. A measure defined by Pincus (1991) and his colleagues (Pincus & Kalman, 1997; Pincus & Singer, 1996), which they call approximate entropy (ApEn), can be computed for any sequence and indicates the degree to which the sequence approximates maximum irregularity. ApEn is maximum when all k-tuples of digits are as nearly equally represented in the sequence as possible. Unpredictability, in the sense that knowledge of a part of a set or sequence does not provide a basis for predicting any other part, follows from the lack of structure; the greater the irregularity or lack of structure, the less the predictability.

A sequence is said to be incompressible if a description of it cannot be given that is shorter than the sequence itself. The sequence of even numbers starting with 2 and ending with 1,000 can be described (to wit the last few words) with fewer characters than would be required to write out the sequence in its entirety; so this sequence is compressible. Another way to make the same point is to say that the sequence is such that it can be generated by a rule or procedure that is shorter than the sequence. The procedure in this case could be: Start with 2; let each successive integer be the preceding integer plus 2; end with 1,000. Following this line of thought, a random sequence is sometimes defined as a sequence the shortest description of which is itself (Chaitin, 1975; Kolmogorov, 1965; Martin-Löf, 1966). The sequence 10101010101010101010 ... is easily compressed, for example, as "alternating 1s and 0's," or as "10 repeated indefinitely." The sequence 00101110100010110101 ... is not so readily compressed. So according to the compressibility criterion, the second sequence might be considered random, but the first clearly would not.

Because any sequence can be represented in binary form, the compressibility criterion is sometimes expressed in terms of the length of a computer program, represented in binary form, that would generate the sequence of interest relative to the length of the sequence itself, also expressed in binary form. The sequence is said to be compressible if the program is shorther than it. Any relatively short binary number is likely to be shorter than a program that would be able to generate it, but if the number is very long and has enough structure to permit it to be produced by a simple rule, the program could be shorter than the number itself.

Some Difficulties

All of these properties have been invoked to characterize randomness, but, although each contributes to the richness of the concept, each also has limitations. The expectation of relatively equal representation, for example, is an *expectation* and neither a necessary nor sufficient condition of randomness. It is not necessary because a random process *can* produce a set in which the items do not appear with anything close to equal frequency. It is not sufficient, because equal representation can be obtained by a process that samples in a highly regular and deterministic way. One would ensure equal representation of the 10 digits in a large number of digits, for example, by selecting the digits in order, from 0 to 9, and repeating this process as many times as necessary to get the size sample desired.

Total irregularity appears to be unattainable. According to an area of mathematics pioneered by Frank Ramsey, any sizable set will contain structured subsets within it, no matter how it was produced (Graham, Rothschild, & Spencer, 1990; Graham & Spencer, 1990).

The idea of minimal description or incompressibility is closely related to the concept of randomness as the absence of pattern or organization. The location of dots on a page would be said to be random if the simplest way to "describe" its "organization" were to present the page itself. If the arrangement admits of a simpler description, or if it could be generated by a procedure that could be described more tersely than the page—that would not have to give the coordinates of the individual dots—it is not random according to this conception.

One attractive consequence of defining randomness in terms of incompressibility (or absence of structure) is that it appears to help avoid the type of perplexing question that was raised in the preceding chapter at the beginning of the discussion of what constitutes an improbable event. By most conceptions of randomness, the outcome of the toss of a fair coin is considered a random event and a sequence of heads and tails produced by 20 tosses would be considered a random sequence. But, as already noted, the probability of getting a specified highly patterned sequence, say, alternating heads and tails, is precisely the same as the probability of getting a specified sequence with no discernible pattern, or a little less than one in a million in both cases.

The compressibility idea seems to permit us to avoid the question of why we should consider one of two equally likely outcomes to be random and the other not, given that they are equally probable. Clearly the one sequence is compressible whereas the other is not. We can state a simple rule for generating the first one, but, it would appear, there is no better way to specify the second one than to write it out explicitly. But consider the following sequence:

$$0011110101110000101 0\ldots$$

This sequence has no apparent structure, but it was generated by a simple rule. The rule was the following one:

> Choose a number x between 0 and 1. Compute the sequence $2x$, 2^2x, 2^3x, ... For each number in this sequence substitute 1 when its fractional part is less than ½ and 0 otherwise.

The value of x that I used was .38. For many choices of x, this rule will produce a sequence of 1s and 0s that will pass most, if not all, of the tests of randomness that one might wish to apply (Kac, 1983). The short—20-digit—sequence that I produced is shorter than the expression of the rule, so giving the rule does not demonstrate its compressibility; but it could be made arbitrarily long by simply applying the rule many times, and any long sequence produced by this procedure can be described in compressed form by stating the rule and specifying the value of x that was used to produce the sequence. And each time the rule is applied with the same value of x, precisely the same random-looking sequence will be produced.

There are other formulas that will generate numbers that appear to be random—that will pass many of the tests that are conventionally used to establish randomness. One such is the recursive formula $x_{n+1} = kx_n(1 - x_n), 0 < x_1 < 1$. For some values of k, x stabilizes at some number; at other values of k, x oscillates successively between two or more numbers. But for certain values of k, x varies haphazardly (Paulos, 1992).

These examples illustrate the asymmetry of the compressibility test for randomness. Anyone who saw a sequence produced by these processes would be hard-pressed to find a way to compress them if he or she did not know the formula that produced them. If one can find a way to compress a sequence, one can say with certainty that, by the criterion of compressibility, the sequence is not random. Inability to find a way to compress a sequence, however, does not guarantee that none exists, so the most one can say in this case is that the sequence has not been shown to be nonrandom.

Random to Whom?

Despite its apparent straightforwardness, the idea of compressibility is problematic in some respects. Determining whether or not a particular arrangement is random in this sense is not always an easy, or even a doable, task. Moreover, if a description is to be useful, it must be comprehensible. Comprehension depends on a store of knowledge that can be applied to the description's interpretation. What is comprehensible to one person may not be to another. The first of the aforementioned rules for generating a random sequence of 1s and 0s will not be comprehensible, for example, to a person who does not know what a fractional part of a number is, or what the use of exponents signifies. Should the knowledge on which the comprehension of a description depends be considered an implicit part of the description, contributing to its length? To Hideaki Tomoyori, who was reported to have been able to recite from memory the first 40,000 digits of π (C. P. Thompson et al., 1991), "the first 1,000 digits of π" would be an effective compression of that string of digits; one who does not have this string stored in memory would have to consult a source that can provide it in order to make use of the "compressed" description.

It seems reasonable to consider information obtained from an external source to be part of the description, from the user's point of view; but then, should we not consider also the critical information stored in Tomoyori's head as part of the description in that case as well? These types of considerations lend support to Shafer's (1993) contention that randomness should be considered not a property of a sequence of numbers, but rather a property of the relation between the numbers and an observer. In particular, Shafer argues that a random-appearing sequence that was produced by a computer program might be nonrandom from the point of view of one who knew of the program but random from that of one who did not: "So the question is not whether or not a given sequence of numbers is truly random; it cannot be random in and of itself. The question is what observer we are talking about. A sequence of numbers generated by a certain program is not random relative to observers who are able to use the program to reproduce them. It may be more or less random relative to observers to whom we deny (or who deny themselves) this ability" (p. 192).

Randomness remains an enigmatic concept. (A very readable discussion of it has been written for a lay audience by Beltrami [1999]). Its usefulness in many contexts, especially the physical and social sciences, is beyond question. A variety of tests of randomness, in addition to the compressibility test, have been developed and are applied in specific instances, but they tend to be indications and not proofs of randomness, and, at best, they represent necessary but not sufficient conditions for considering something to be random. That is, if a set of numbers fails to pass a given test, one may conclude that it is not random; but if it passes, the most one can conclude is that it *may* be random, the test did not rule out the possibility. If one wants to have a high degree of confidence that one has a means of producing random sequences, the best one can hope for is that the process will yield sequences that are not shown, by some of the more demanding tests that can be applied, to be nonrandom; for most practical purposes this appears to be more than enough.

RANDOM PROCESSES AND RANDOM PRODUCTS

Implicit in much of the foregoing discussion has been a distinction between a random *process* and a random *product,* a distinction that has been made explicit by several writers (e.g., Falk & Konold, 1997; Zabell, 1988). Conceptually the difference between a process and a product is clear: but how to operationalize it in practice is not obvious. Some investigators argue that randomness is more appropriately applied, as a descriptor, to a process than to a product (Lopes, 1982; Wagenaar, 1991), or take the position that a random product should be defined as the output of a random process (Pollatsek, 1991).

Disagreement on the question of whether randomness is better thought of as a property of a process or as a property of the outcome of a process is not new and not likely to be resolved to everyone's satisfaction soon (Bennett, 1998).

A random process may be defined as one that selects items from a set in such a way that every item in the set has the same probability as every other item of being selected on each trial. A random product could be defined as any product produced by a random process; alternatively it could be defined in terms of characteristics, like incompressibility, that make no reference to the way it was produced. According to the process-based definition of a random product, HHHHHHHHHHHHHHHHHHHH, representing the tossing of 20 heads in a row with a fair coin, would be considered a random sequence; according to a definition based on incompressibility, such a sequence would be considered distinctly nonrandom.

In other words, random processes can produce products that are nonrandom, as judged by criteria that are not process based. And, as we have seen, deterministic processes can produce products that are random, as judged by the same criteria. Add to this the fact that the way one usually decides whether a process is random is by checking the products it produces, and it is easy to see why the concept of randomness is more than a little elusive. The usefulness of the concept, despite its shaky conceptual foundation, is remarkable. From a practical perspective, it is not necessary to have a definition of randomness on which everyone agrees; it suffices to be able to produce sets that are random in a sense and to a degree that suits the purpose for which they are to be used, and what is satisfactory for one purpose may not be for another.

TESTING FOR RANDOMNESS

How should one test for randomness? One answer to this question that has been proposed is that there is no way to demonstrate, beyond all doubt, that the output of a given process is truly random. Horwich (1982), for example, argues that no finite sequence can be shown to be entirely representative of the long-term output of a stochastic process; conversely, one can argue that any finite sequence could be the output of a stochastic process, because, according to the theory, every sequence that has nonzero probability of occurring will be produced by a random process if only it runs long enough.

It is sometimes held that establishing that a nonrandom sequence is not random is more straightforward than establishing that a random sequence is random (Wagenaar, 1972), but this is debatable. Given that a random process can, and will, produce sequences that have as much structure as one pleases, one can never say with certainty by inspecting a sequence that it is not random if one defines a random sequence as a sequence produced by a random process.

Randomness as a Contingent Property

Twenty tosses of a coin that produced 20 heads would, by just about any test, be judged to be nonrandom, but 20 consecutive heads embedded in a sequence of a few million tosses would not be considered evidence that the entire sequence was nonrandom; to the contrary, the larger sample would be suspect if it had no runs of this length. It follows from this type of consideration that it is possible to have many sequences that pass conventional tests of randomness that, when joined together, yield a much longer sequence that fails because it does not contain sufficiently long runs (Ayton, Hunt, & G. Wright, 1991a).

Suppose we invoke the idea of compressibility. Devlin (2000b) gives the following sequence as one he generated by rolling a fair die 42 times: 1, 3, 3, 6, 5, 6, 6, 4, 3, 5, 2, 4, 6, 1, 5, 2, 1, 2, 5, 6, 1, 1, 4, 3, 5, 4, 1, 4, 4, 1, 5, 1, 3, 3, 2, 2, 6, 2, 6, 4, 3, 5. This sequence probably satisfies the criterion of noncompressibility; it is hard to imagine producing a description of it that is shorter than the sequence itself. So, according to this criterion, we would call the sequence random. Devlin goes on to argue:

> But suppose I doctored the list a bit, changing the first 1 to a 2? Not only would such a sequence be the product of design (because of my tampering with the original result), but it would also carry the mark of design, inasmuch as each number from one through six would occur exactly seven times. Still, there seems to be no way to specify the new list of numbers that is shorter than the sequence itself.... Hence, according to Kolmogorov's definition, the sequence is random, despite the element of design in its generation. (p. 15)

(Devlin's treatment of equal representation as evidence of design may seem contrary to the treatment a few pages back of equal representation as evidence of randomness. But note that the earlier reference said *approximately* equal percentages after a *large number* of selections. The qualifications are important; when actual relative frequencies match expected relative frequencies too closely, especially in small samples, suspicion of tampering is aroused.)

Consider the distribution of points on the area shown in Figure 2.1. Is this distribution random? What does it mean for the distribution of points within an area to be random? One thing it might mean is that each point is equally likely to appear anywhere within the area. This definition implies that if the area were subdivided into subareas of equal size, the numbers of points falling in the different subareas should not differ from each other more than would be expected by chance. So if we wanted to determine whether the distribution of points in the figure is random, we might do a statistical test to see if the numbers of points within equal-size subareas differ more than would be expected by chance. The chi-square test was designed for just such purposes.

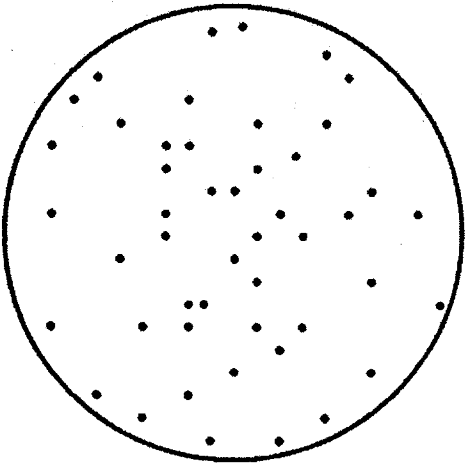


FIG. 2.1. Are the dots in this figure randomly distributed?

There are, of course many ways in which a circular area can be divided into subareas of equal size. Figure 2.2 shows a few of them. Which of these divisions or other possibilities should be used to test whether the distribution of points in Fig. 2.1 is random? Presumably it should make no difference. If a test using one division showed the distribution to be random and one with a different division showed it to be nonrandom, we would wonder either about our concept of randomness or about the appropriateness of our test for it.

If we divide the circle into four quadrants, a chi-square test is likely to come out one way if the division is made horizontally and vertically than if it is made with the two diagonals (bottom two divisions in Fig. 2.2). In the first case, the test will not force rejection of the null hypothesis and will therefore lead us to conclude that the assumption of a random distribution is tenable. In the second case, the test will make us reject the hypothesis of no difference and conclude that the distribution is not random. So if we want to believe that the distribution is random, we should perform the test with the first division, and if we prefer to believe that it is not random, we should do it with the other one.

It is generally agreed by statisticians that this way of proceeding is cheating. In situations of this sort, one is likely to be able to find a way of partitioning the world so as to get a desired outcome if one has the opportunity to scrutinize the data before deciding how to design the test. This is one of the reasons that many statisticians insist that the hypotheses that are to be tested by experimentation and the statistical procedures that are to be used to test those hypotheses be specified in detail before the experimental data are collected.

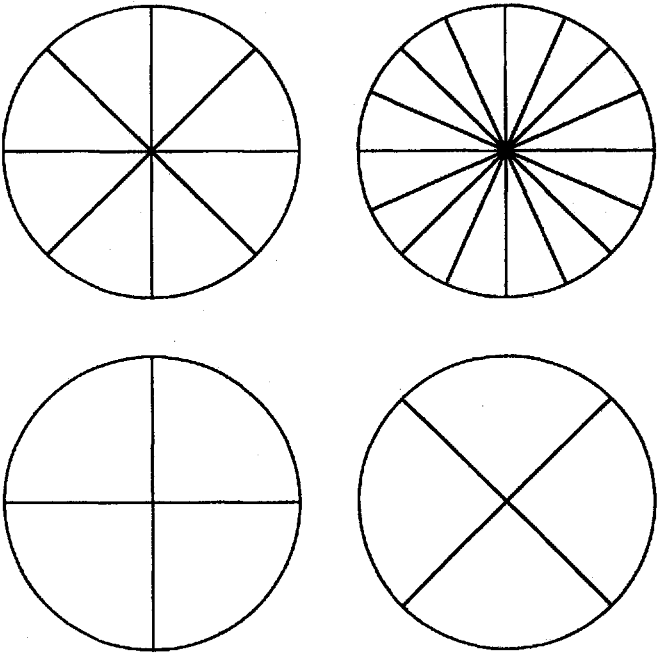


FIG. 2.2. Possible ways to partition the area of a circle to test for the randomness of the distribution of points in the circle.

The decision as to what would be an appropriate test for randomness, even when made before the collection of data, must be guided by the nature of the hypothesized nonrandom phenomenon of interest. Suppose that we wish to know whether some process that is about to distribute some points in a circular area will distribute them at random. We might agree, before the fact, to divide the area into four equal pie-shaped quadrants and to use chi-square to test the hypothesis that the number of points ending up in each of the various quadrants does not differ more than would expected by chance. This seems like an unbiased approach, provided we specify the boundaries between the quadrants in advance of seeing any data.

But suppose that points represent the locations of darts thrown at a bull's-eye within a circular target and the subject of interest is the accuracy of the dart thrower. Specifically, imagine that we wish to decide which of the two hypotheses is the more tenable: (a) the points are distributed randomly on the target or (b) the points are more clustered around the bull's-eye than would be expected if they were randomly placed. A test based on division of the circle into pie-shaped quadrants will tell us something about randomness with respect to angular dispersion but nothing about whether the clustering of the points around the bull's-eye is greater than would be expected by chance. Clearly the test should take into account the location of the darts relative to the bull's-eye. But there is still a choice to make.

One way to proceed would be to draw a circle around the bull's-eye so as to divide the target into two subsections of equal area, as shown in Fig. 2.3, and then use an appropriate statistical test of whether there are more dart points in the inner area than in the outer one and the difference is significantly greater than would be expected by chance. Alternatively, one might represent each dart location by its Euclidean distance from the center of the target area and do an appropriate statistical test of whether the dart locations that are less than half the length of the radius from the center outnumber those that are more than half the length of the radius from the center and whether the difference between these numbers is greater than would be expected by chance.

Either of these approaches to deciding whether the dart locations are more clustered around the center of the target than would be expected by chance seems to make sense. Unfortunately they are not equivalent. The area of the inner circle shown in Fig. 2.3 is half that of the target, and thus equal to the area of

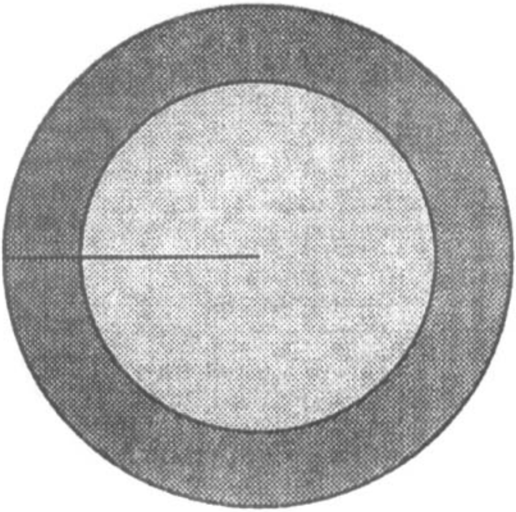


FIG. 2.3. The area of a circle partitioned as it might be to determine whether points within it representing the locations of darts thrown at a bull's-eye are randomly dispersed over the entire area. The area of the inner circle is equal to half the area of the larger one, that is, equal to the area of the larger circle minus the area of the smaller one.

the torus surrounding it. But all the points that are less than half the length of the radius of the target from the center of the target define an area, as shown in Fig. 2.4, that is only one quarter the area of the whole target, or one third of the area of the torus surrounding it. It would be possible to have a distribution of dart locations that would be considered clustered around the center of the target according to the first criterion, represented by Fig. 2.3, but that would be considered a chance distribution by the second criterion, represented by Fig. 2.4.

What appears to be a difficulty follows from the nonlinear relationship between a circle's area and its radius, $A = \pi r^2$, and in particular the fact that $\frac{\pi r^2}{2} \neq \pi \left(\frac{r}{2}\right)^2$. When the areas of two circles are in the ratio 2:1 their respective radii are in the (approximate) ratio 1:4.1; when the ratio of their radii are in the ratio 2:1 their respective areas are in the ratio 4:1. So which of the two approaches considered is the appropriate one to use to test the question of whether or not the dart locations are random with respect to the center of the target? The answer is one cannot say without being more precise about what one wants to mean by random in this context. Imagine being given the task of distributing points on the target area in such a way that their locations would be

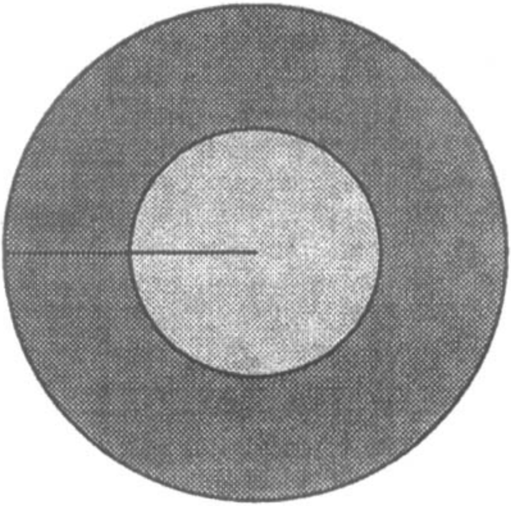


FIG. 2.4. The ratio of the areas of the outer and inner circles is 4:1; the ratio of the radii of the outer and inner circle is 2:1.

random with respect to the center. One way to proceed would be to superimpose a grid on the target and select the *x*, *y* coordinates of each point using a pair of numbers from a table of random numbers. This would produce a distribution of points that is expected to be roughly the same in any two regions of equal area, and, given this operational definition of randomness, the first approach mentioned earlier would be appropriate.

Alternatively, one might determine the random points, again by drawing pairs of numbers from a random number table, but this time using the first number of each pair to designate the distance of a point from the center and the second one to determine the orientation of a radius drawn through the point with respect to the 360 degrees of the circle. This would produce a distribution that is uniform with respect to distance from the center, but with a greater density of points for a region closer to the center than for a region of equal area farther from the center, and given this operational definition of randomness, the second approach mentioned earlier would be appropriate.

One may have good reason to prefer one of these definitions of randomness over the other for a specific purpose, but neither can be said to be correct in an absolute sense. The important point is that failure to be specific about what one means by randomness in particular contexts can be problematic. We will encounter this fact again in the chapter on paradoxes and dilemmas, in particular in a discussion of Bertrand's paradox.

Another point regarding testing for randomness that is implicit in the preceding comments is that variables can be random in some respects and structured in others. One might say, for example, that the distribution of points in a circle is random with respect to the horizontal dimension, but nonrandom with respect to the vertical. Or we might describe a distribution as random with respect to angular displacement from the vertical but nonrandom with respect to distance from the center.

Predicted Versus Discovered Randomness

The rule that hypotheses and statistical testing processes should be specified before the collection of experimental data seems a good one to apply whenever experimentation is to be done. This does not mean, however, that what appears to be structure that was not predicted or anticipated should be ignored. When one discovers by observation that there is a way of looking at a data set that makes it appear nonrandom, as, for example, when one observes that one way of partitioning a space is likely to produce statistical evidence of structure whereas another partitioning is not, this can be a useful hint for further experimental exploration. Such evidence of structure must be considered tentative, and should never be reported without a clear explanation of the role that selection played in revealing it, but it can be of considerable interest, nonetheless.

Gilovich (1991) discusses an interesting case of after-the-fact assessment of the randomness of a distribution of points over an area based on studies by Clarke (1946) and D. Johnson (1981). The question of interest was whether the V-1 bombs dropped on central London by the German airforce during World War II were randomly distributed over the entire area. Visual inspection of the distribution (see Gilovich, p. 20) suggests a greater-than-chance concentration in the northwest and southeast quadrants. And indeed, a chi-square test would permit rejection of the hypothesis of equal distribution over the four quadrants with a high degree of confidence. But is the performance of such a test appropriate, given the absence of an a priori reason for partitioning the area into quadrants by dividing it north–south and east–west, and the fact that interest in the possibility of a nonrandom distribution was stimulated by inspection of the data? Gilovich says unequivocally no, on the grounds that "with *hindsight* it is always possible to spot the most anomalous features of the data and build a favorable statistical analysis around them" (p. 21). The reason for partitioning the map into four rectangular quadrants, instead of dividing it into four sectors with two bisecting diagonals, say, is to maximize the sensitivity of the test to the structure that we think we see by inspecting the figure.

A more extreme case of selectivity in testing for randomness is seen in the following illustration. Consider a set of points distributed as shown in Fig. 2.5. It is easy to imagine a partitioning of this space for which a chi-square test would provide evidence of randomness. If one divides the area into 16 equal squares as shown in Fig. 2.6, for example, one gets a chi-square of 11 that, with 15 degrees of freedom, yields a p of about .75, so one cannot reject the hypothesis of a chance distribution. But in inspecting the original distribution, one may notice that there appear to be more points in the upper right and lower left quadrants than in the upper left and lower right. If one divides the total area into four equal-size squares, as in Fig. 2.7, and does a chi-square test, one gets a chi-square of about 8.6 that, with 3 degrees of freedom, yields a p of less than .05, so one can now reject the hypothesis of random distribution.

Is one entitled to conclude from this second analysis that the points are not randomly distributed within the entire area? Many statisticians would say definitely not. I want to argue that the question requires a qualified answer. One can certainly say that the distribution of points in the space does not pass the particular (second) test of randomness that was used. Because of the way in which the test was applied, one does not have a good basis for drawing any firm conclusions; however, the observed pattern of results raises a question as to

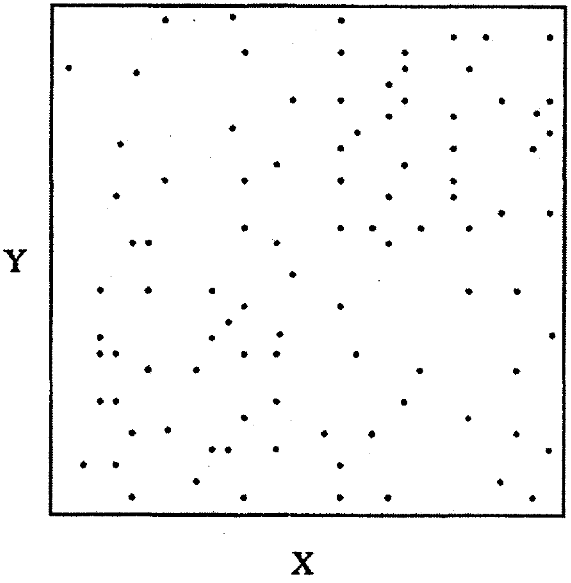


FIG. 2.5. Is the distribution of points in the square random?

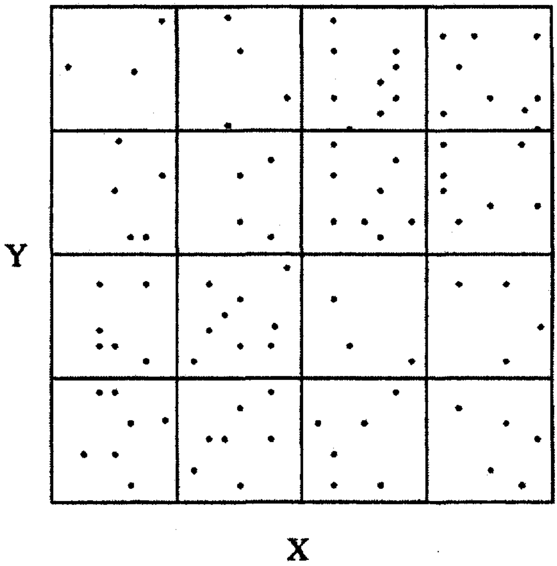


FIG. 2.6. A chi-square test with this partitioning of the square supports the hypothesis that the points are distributed randomly.

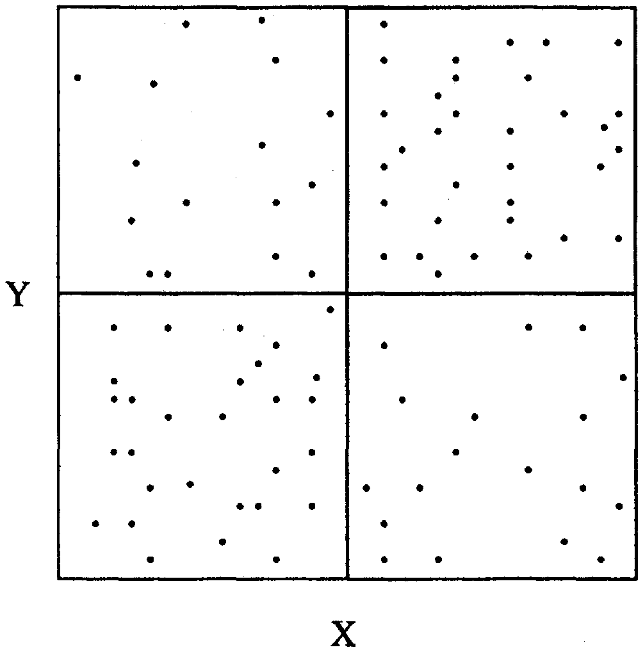


FIG. 2.7. A chi-square test with this partitioning of the square supports the hypothesis that the points are not distributed randomly. There appears to be greater clustering in the lower left and upper right quadrants than in the upper left and lower right ones.

whether there may be some causal reason why the points are clustered more along the major diagonal (lower left to upper right) than along the minor one (upper left to lower right). It would be foolish to ignore this possible clue to structure; it provides a reason for making an effort to check more data with this hypothesis in mind.

Although the illustrations in this discussion have been spatial, the issue of selectivity is relevant to the problem of distinguishing between randomness and nonrandomness more generally. Imagine a long sequence of 1s and 0s of the kind that might be generated by many tosses of a coin. If one were to scan such a sequence looking for a subsequence of length 20, say, that appeared to be nonrandom, and that would be shown to be nonrandom by the application of some standard statistical test, one would be very likely to be able to find one. This method of finding structure would be rejected by statistical sophisticates, of course. We should note, however, that in principle the approach is not unlike the practice of repeating a failed experiment until it yields the desired statistical result.

Every so often, there is a report in the news media of a neighborhood or town that has become alarmed because there appears to be a higher-than-chance incidence of some disease (usually cancer) among its residents. From the perspective of a resident, this is an understandable cause for concern and for an effort to determine whether there is a causal explanation for the unusually high rate of the condition. However, from the perspective of a demographer, the finding that among the many thousands of towns and neighborhoods in the country, a few have an unusually high rate of cancer would not be surprising; this is to be expected on a purely statistical basis, just as an occasional run of 10 consecutive heads is to be expected if one tosses a fair coin a sufficiently large number of times. One cannot blame the resident of a high-incidence town for looking for something unusual about the area that would account for the rate, but one should not be surprised, either, if there is nothing unusual to be found.

As already noted, any sizable set will contain structured subsets within it, no matter how it was produced; so, if one looks for structure, one is very likely to find it, even in data that have been produced by a random process. This being said, it would be hard to deny that the adventitious discovery of structure has played an important role in science and is a critical aspect of effective thinking more generally as well. The moral of the fact that structure may be found even in products of random processes is that discovered structure should be taken not as strong evidence of causal forces at work but as a stimulus for further investigation—controlled experimentation when feasible—aimed at providing independent evidence as to the tenability of the hypothesis that the observed structure is real and not simply the result of having selected a low-frequency outcome of a random process.

However, not all the hypotheses that one might want to entertain regarding randomness lend themselves to testing via controlled experimentation. Astronomers, for example, are interested in the question of whether stars and galaxies are distributed randomly throughout the universe. They are quite convinced that, except on the largest scale, the degree of clustering is greater than would be expected by chance, and this conclusion has been drawn on the basis of after-the-fact observation. No one proposed a hypothesis about the distribution of stars and specified a statistical process for testing it before the actual distribution of stars was observed. It was by looking at the actual distribution that astronomers got the idea that it is nonrandom in specific ways. When data suggesting structure are obtained from observation and they are not subject to experimental corroboration, as in the case of the distribution of stars, perhaps the best that can be done is to consider the aggregate weight of all the data that can be brought to bear on the question of interest.

THE PRODUCTION AND PERCEPTION OF RANDOMNESS

> Charles Dickens is said to have refused, late one December, to travel by train because the annual quota of railroad accidents in Britain had not yet been filled that year.
>
> Nature abhors a vacuum, human nature abhors chaos. Show us randomness and we will find order, pattern, clusters, and streaks. (Myers, 2002, p. 134)

In view of the difficulty that statisticians have had in agreeing on the nature of randomness, we should not be surprised if lay people often have an imperfect understanding of the concept. Many experiments have been done to determine how good people are at producing or identifying random sequences or sets (Nickerson, 2002). The general conclusion that the results of these experiments in the aggregate seem to support is that people are not very good at these tasks—that they find it hard to generate random sets on request and to distinguish between those that have been produced by random processes and those that have not. It appears that people often consider sequences that have been generated by random processes to be nonrandom and they see contingencies where they do not exist (J. Cohen, 1972; Wagenaar, 1972). Studies of the perception of covariation, for example, some of which have been motivated by an interest in superstitious behavior, have yielded evidence that people sometimes impute contingency relationships between variables that are independent (Catania & Cutts, 1963; Hake & Hyman, 1953; J. C. Wright, 1962). The gambler's fallacy, the negative-recency effect, the law of small numbers, and a variety of other concepts attest to the prevalence of the belief among researchers that people find it hard to distinguish consistently between random and nonrandom sets and to generate random ones on request.

The tenability of the conclusion that people are poor producers and perceivers of randomness has been challenged on the ground that much of the work has been predicated on the assumption that there exist valid objective criteria with which to judge the adequacy of subjective conceptions of randomness and nonrandomness (Ayton, Hunt, & G. Wright, 1991a, 1991b). Ayton et al. (1991a) argue not only that this assumption is false but that in many cases "psychologists have set their subjects the task of generating or recognizing random sequences, without explicitly defining what sort of sequence would count as patterned, and therefore nonrandom, and then do not demur from passing judgment on the adequacy of the performance of the task" (p. 224). They suggest that often, in telling people to attempt to produce sequences that have certain properties (appear to be jumbled or orderless) or that fail to have others (structure or regularity), they are, in effect, instructing them to produce sequences that have properties (e.g., local representativeness, no long runs, etc.) that are later taken as evidence of nonrandomness.

Other investigators have argued that the controversial nature of the concept of randomness and lack of agreement among experts as to how it should be defined make it difficult to assess the ability of nonexperts to recognize or produce randomness, because the standard against which their performance should be judged is unclear (Lopes, 1982). If experts do not agree as to what constitutes randomness, it is not clear what we should assume that nonexperts take the term to mean. One suspects that there are large individual differences in this regard and that many of the conceptions are imprecise. Evidence that people produce sequences that are more nearly random by conventional tests when asked to make them unpredictable than when asked to make them random (Finke, 1984) suggests that randomness and unpredictability may not be equivalent in many minds.

A review of experimentation on people's ability to produce or perceive randomness (Nickerson, 2002) revealed the importance of task instructions and the difficulty of interpreting results when instructions are vague or ambiguous, as they often have been. The results of many experiments have been taken as evidence that people are not good at generating or recognizing randomness, but, although the conclusion might be correct, much of the experimental support that has been advanced for it is weak because of the ambiguities involved.

COMMON MISUNDERSTANDINGS INVOLVING EVENT INDEPENDENCE

The idea of randomness is closely associated with that of event independence. A random sequence of coin tosses, for example, is one in which the probability of a head on each toss is .5, or, in other terms, the outcome of each toss is independent of the outcomes of the preceding tosses. Some misunderstandings of randomness seem to have their basis in the imputation of contingencies among independent events, or lack of acceptance of the idea of event independence.

The "Gambler's Fallacy"

Perhaps the best known case of an imputed contingency is the "gambler's fallacy," according to one form of which a run of successive occurrences of one type of random event (e.g., a run of heads in coin tossing) will make an additional occurrence of that event appear to be less likely. Another, but closely related, form is the belief that when a sample of ongoing random events shows a "deficit" of one type of event, the probability of the imminent occurrence of that event is increased. If, for example, heads has outnumbered tails in a series of coin tosses, the probability of the occurrence of tails is assumed to be increased until the balance is restored. Laplace (1814/1951) gives an amusing example of this form of the bias: "I have seen men, ardently desirous of having a son, who could learn only with anxiety of the births of boys in the month when they expected to become fathers. Imagining that the ratio of these births to those of girls ought to be the same at the end of each month, they judged that the boys already born would render more probable the births next of girls" (p. 162).

Closely related to the gambler's fallacy is the "negative-recency" effect, which refers to a bias against repetition evidenced by people attempting to generate or identify random sequences (Bar-Hillel & Wagenaar, 1993; Wagenaar, 1970). A spatial analog to the negative-recency effect has been observed when people make presumably random selections from a list and avoid selecting adjacent items (H. C. A. Dale, 1960). A negative-recency effect appears to be operative in the attitudes many people express regarding the occurrence of natural disasters that are generally assumed to be randomly timed; many people seem to believe that the occurrence of a natural disaster (earthquake, flood, tornado) more or less guarantees that such an event will not recur for a relatively long time (Burton, Kates, & White, 1978).

The gambler's fallacy can take more complicated forms as, for example, when in a two-alternative prediction task people tend to predict the more frequent event after one occurrence of the less frequent event and to predict the less frequent event after two consecutive occurrences of the more frequent one (Jarvik, 1951). Study of the gambler's fallacy is complicated by the fact that people sometimes make predictions that are consistent with the assumption that they believe that independent events are contingent even when they indicate, when asked, that they believe them to be independent. Bar-Hillel and Budescu (1995) cite this finding, reported in an unpublished manuscript by Gold and Hester (1989), as evidence of the need for caution in inferring what people believe about probabilities from their predictions of the outcomes of probabilistic events.

When attempting to predict the outcomes of binary events (win-loss of a bet on a coin toss, win-loss of a football game) people sometimes use information regarding previous outcomes differently, depending on what they assume about how the outcomes are determined. When outcomes are assumed to be determined by chance, as in the case of bets on coin tosses, previous wins are likely to lead to prediction of a loss; when they are assumed to be determined causally, as in the case of football games, previous wins are likely to lead to prediction of a win. The first case is an example of the gambler's fallacy. The second has sometimes been viewed as an example of a complementary misconception of chance within the world of sports, which is illustrated by the "hot-hand" phenomenon in basketball: It is commonly believed among basketball players and fans that players sometimes experience an unusual streak of successful shots; on such occasions they are said to be hot, or to have a hot hand (Gilovich, Vallone, & Tversky, 1985).

The "Hot Hand" and Streaky Performance

> In a long series of events of the same kind the single chances of hazard ought sometimes to offer the singular veins of good luck or bad luck which the majority of players do not fail to attribute to a kind of fatality. (Laplace, 1814/1951, p. 164)

Some investigators have argued that both the gambler's fallacy and the hot-hand belief rest on a misunderstanding of randomness, and, in particular, on unawareness of the frequency with which runs of moderate length are likely to occur in randomly generated sequences; runs in small samples are more likely than they are generally believed to be (Wagenaar, 1972). Gilovich (1991) explains the hot-hand belief this way: "Players and fans are not mistaken in what they see: Basketball players do shoot in streaks. But the length and frequency of such streaks do not exceed the laws of chance and thus do not warrant an explanation involving factors like confidence and relaxation that comprise the mythical hot hand" (p. 16). The mistake that players and fans who believe in the hot-hand phenomenon make lies, according to this view, not in what they see but in how they interpret what they see. The claim that the length and frequency of streaks of successful shots do not exceed the laws of chance is based on an analysis of shooting records of professional players by Gilovich et al. (1985).

Gilovich et al. (1985) note that they do not attempt in their analysis to capture all that people might mean by "the hot hand" or "streak shooting," but they argue that the common use of the terms implies that the sequences of hits and misses of basketball shots should differ from sequences of heads and tails produced in coin tossing in two ways: (a) the probability of a hit should be greater following a hit than following a miss, and (b) the number of streaks of successive hits should be greater than the number produced by a chance process with a constant hit rate. The data presented by Gilovich et al. constitute evidence against the idea that the probability of a hit is greater, on the average, immediately following one or a sequence of other hits than following one or a sequence of misses.

Some dyed-in-the wool sports fans who are aware of the data presented by Gilovich and colleagues are reluctant to accept them as compelling evidence that there is no such thing as a hot hand (Hooke, 1989; Larkey, R. A. Smith, & Kadane, 1989). I confess to being among them, and although I realize that that reluctance may simply bear out Gilovich's (1991) observation that belief in the hot hand within the basketball world is very strong and not easily changed by the presentation of counterindicative evidence of the sort that he and his colleagues have produced, I will try to say why I find their data unconvincing. I do not mean to deny that observers often see evidence of conditional dependencies where none exists. This was demonstrated convincingly by Tversky and Gilovich's (1989a) finding that believers in the hot-hand phenomenon are sometimes convinced they see streak shooting in randomly generated data. But from the fact that people sometimes see dependencies that do not exist, it does not follow that there are never such dependencies to be seen.

The hot-hand or streaky performance can mean at least two things: (a) that the outcomes of successive trials are not independent (e.g., that the probability of success on trial *n*, given success on trial *n* – 1, is greater than would be predicted from a knowledge of the overall probability of success and the assumption of trial independence); (b) that a player's short-term steady-state probability of success fluctuates over time more than is consistent with the assumption of random variation around a stable long-term probability—that individual players experience limited periods of time during which they play significantly better than they do on average. Gilovich et al. recognize the latter conception in noting that references to hot-hand or streak-shooting phenomena "express the belief that the performance of a player during a particular period is significantly better than expected on the basis of the player's overall record" (p. 295). In my view, the data they present do not constitute strong evidence against the occurrence of this type of hot-hand phenomenon.

According to this conception, a hot (or cold) hand would reveal itself, not in the difference between the probability of a hit conditional on a preceding hit and the probability of a hit conditional on a preceding miss, but in a short-term increase (or decrease) in a player's hit rate. To be specific, when a player who has a long-term-average hit rate of .5 plays a game in which his hit rate is .7, one might say that during that game he had a hot hand. Conversely, when the same player plays a game with a hit rate of .3, one might say that his hand, during that game, was cold. Of course, short-term fluctuations in hit rate would be expected strictly on the basis of chance, just as the proportion of heads in short sequences of coin tosses would be expected to deviate from precisely .5. However, if the probability of the outcome is really constant, and not varying over time, one expects the extent to which the relative frequencies of outcomes in small samples will deviate from that probability to be limited and within specifiable bounds.

Gilovich et al. (1985) addressed the possibility that the short-term hit rate varies from the long-term average more than would be expected by chance, assuming a fixed underlying probability, in two ways. They compared two estimates of the standard error of individual players' per-game shooting percentage, one based on the player's shooting percentages for each game and the other from his overall shooting percentage across games. They also looked for evidence of more sequences of successive hits than would be expected by chance if the hit rate were constant over time. The rationale for the latter analysis was the assumption that if a player is occasionally hot, his performance record ought to reveal a greater number of hit streaks than would be expected by chance. Both analyses failed to yield evidence of a nonchance effect. This result justifies the claim that no evidence for the reality of the hot hand was found, but not, unless one wishes to assert that the null hypothesis has been proved, to claim that it has been shown not to exist.

Larkey et al. (1989) argue that the analyses from which Gilovich et al. (1985) concluded that streak shooting is illusory ignore the effects on performance of game context and how a player's shooting interacts with the activities of the other players. On the basis of an analysis of the shooting records of 18 star players during the 1987–1988 NBA season that was designed, in their view, to take game context into account, these authors concluded that streak shooting does occur and offered Vincent ("the Microwave") Johnson as a *bona fide* streak shooter. Although I agree with Larkey, Smith, and Kadane's point that any analysis that purports to demonstrate the existence or nonexistence of streak shooting should take game context into account, their own analysis was criticized, effectively in my view, by Tversky and Gilovich (1989b).

Gilovich et al. (1985) make it clear how randomly produced runs can easily be misperceived as hot-hand effects. And the results of their analyses certainly give believers in the hot-hand phenomenon something to think about. As Hooke (1989) has argued, however, they do not justify the conclusion that no such thing as a hot hand exists, which is what designating belief in the hot hand as erroneous and as based on a cognitive illusion seems to imply. It may be the case that the belief is erroneous and that the basis for it is totally illusory, but this has yet to be demonstrated. There is a difference between saying that no one has produced compelling evidence that it exists and saying that it does not exist. And although some of the people who have written on this subject recognize this difference, not all do. S. J. Gould (1989), for example, citing the data presented by Tversky and Gilovich (1989a), says flatly of the hot hand, "no such phenomenon exists" (p. 12). One might argue that in the absence of compelling evidence of its existence the assumption of its nonexistence should be made on the principle of parsimony, but we should distinguish an assumption that is made in the interest of parsimony and that is not ruled out by evidence from a conclusion that is forced by evidence.

Hooke (1989) makes what seems to me a critical point in contending that while we know what should happen if only chance is involved, we do not have a good idea of exactly what to expect if streak shooting actually exists; in his words, "we don't have a well-formulated hypothesis to test against the null hypothesis" (p. 36). Also, the null-hypothesis testing that has been applied to hot-hand data has been primarily, if not exclusively, rejection-support (RS) testing as contrasted with acceptance-support (AS) testing. Normally RS testing is appropriate when the null hypothesis represents what the experimenter does not believe and expects to be able to reject, and rejection is to be taken as evidence supportive of the experimenter's theoretical position. (See also Meehl's [1967, 1990, 1997] distinction between strong and weak uses of statistical significance tests.) AS testing is appropriate when the null hypothesis represents what the experimenter believes and *acceptance* of it would be taken as support for the experimenter's view (Binder, 1963; Steiger & Fouladi, 1997). In RS testing, the decision criterion is biased against Type I error (rejection of the null hypothesis when it is true), while in AS testing, the criterion is biased against Type II error (acceptance of the null hypothesis when it is false). Using RS testing is questionable when one's intent is to show the null hypothesis to be true (i.e., to show the hypothesis that there is such a thing as a hot hand to be false).

One basis for reluctance to accept the conclusion that the hot hand is illusory in the absence of evidence that forces one to this conclusion is the implausibility—to me at least—of the implied hypothesis that an athlete's performance is precisely constant over time. It is too easy to think of plausible reasons why a player's (shooter's) performance (hit probability) might, unlike the fair coin, differ from one time to another. These include general health, fatigue, mental state, playing arena, teammate combinations, and opposing players. One might argue that the effects of such variables, especially in combination, can be assumed to be random, and for a large enough sample, perhaps they can; but this does not rule out the possibility that a player's hit rate might be better than his average when playing, uninjured and well rested, on his home court, with his favorite play maker on the floor, and against a weak defender.

I suspect that the truth about the hot hand lies somewhere between the extreme view that sees it whenever a player's short-term success rate is noticeably above his long-term average or he makes several baskets in a row and the other extreme position according to which it never occurs. Believers in the hot-hand phenomenon may be too quick to attribute any impressive sample of shooting performance to it; nonbelievers may be too quick to rule out the possibility that shooters may sometimes be genuinely better than they typically are.

Evidence that performance in sporting events sometimes is streaky comes from a study by Gilden and Gray Wilson (1995). Using a "runs *z* score" as "a measure of outcome clustering that is independent of both sequence length and hit rate," these investigators found evidence of streaky performance in both golf putting and dart throwing. In both cases, the probability of streaks appeared to be a nonlinear function of task difficulty, being greatest for intermediate difficulty levels and smaller or (in the case of dart throwing) not different from chance with very easy and very difficult targets.

Waldrop (1995) has recently reported an analysis of some of the data considered by Tversky and Gilovich (1989a) that suggests that the question of whether the hot hand is an illusion may be more difficult to answer than one might have assumed. The data on which Waldrop focused were the 1980–1981 and 1981–1982 free-throw shooting records of nine regulars of the Boston Celtics. The question of interest was whether there was any evidence in these records of streak shooting in the first sense mentioned previously—the probability of a hit on the second of two shots being higher following a hit than following a miss. Tversky and Gilovich had concluded from their analysis of the same data that there was no evidence that the outcome of the second shot depended on that of the first.

Waldrop (1995) notes that one comes to different conclusions about the dependencies if one examines the data on a player-by-player basis than if one looks at them in the aggregate. Roughly half of the players did better on the second shot if they missed the first one and the other half did the reverse, but none of the differences is very impressive from a statistical point of view. When the data are pooled over players, however, the probability that the second shot was a hit is significantly greater if the first shot was a hit than if it was a miss. Waldrop argues that basketball fans are more likely to have a representation in memory that corresponds to the aggregate contingency table than a representation of a table for each individual player and that, therefore, their belief that the second of two shots is more likely to be a hit if the first one was a hit should not be considered unfounded, at least in the case of free throws. This does not justify the belief with respect to individual players, but does so with respect to a team as a whole. (Waldrop also makes a statistical argument that several players did better on their second shot than on their first one, and argues that while this outcome does not support the hot-hand hypothesis, it does count against the hypothesis of independence.)

Waldrop's (1995) analysis involves a paradox—Simpson's paradox—that is discussed in chapter 6. It is most readily illustrated in this context by focusing, as Waldrop did, on the performance of Larry Bird and Rick Robey, the best and worst free-throw shooters, respectively, on the team. As shown in Table 2.1 both Bird and Robey were somewhat more likely to make the second of two free throws if the first was a miss than if it was a hit, but when their data are combined, the opposite relationship holds. The reversal is accounted for in this case by the fact that Bird has more impact on the combined table than does Robey by virtue of taking more free throws and making a larger percentage of them. It turns out that the differential weightings of the individual players' results has a similar effect when the data are combined for the whole team.

TABLE 2.1

Frequencies of Pairs of Free Throws of Bird and Robey Individually and Combined

	Bird			*Robey*			*Combined*		
	—H	*—M*	*Σ*	*—H*	*—M*	*Σ*	*—H*	*—M*	*Σ*
H—	251	34	285	54	37	91	305	71	376
M—	48	5	53	49	31	80	97	36	133
S	299	39	338	103	68	171	402	107	509
P(H|H)	251/285 = .881			54/91 = .593			305/376 = .811		
P(H|M)	48/53 = .906			49/80 = .613			97/133 = .729		

My sense is that the tale of the hot hand has not yet been told in its entirety. There can be no doubt that people see structure in events that have been generated by random processes and sometimes misinterpret that structure as evidence of nonrandom effects. That *all* the apparent structure that is interpreted as evidence of a hot hand or similar phenomenon is the result of random processes is much less clear.

Nonaging Processes

Ernest Rutherford discovered radioactivity and introduced the concept of the half-life of a radioactive atom at the turn of the 20th century. According to his theory, the probability that an atom of a radioactive element will decay instantaneously differs for different elements but, for any given element, is constant over time. The atom does not age; the probability that it will decay in the next instant is constant and independent of how long it has been in existence. This means that a fixed proportion of the remaining atoms of a substance will decay in a fixed interval of time. It follows that the *number* of atoms of that substance remaining will decrease exponentially with time, the rate of decrease depending on the substance involved.

A nonaging process is one for which the remaining "life expectancy," like that of a radioactive atom, is independent of its current age. If its life expectancy at age 1 was 17, its remaining life expectancy at 7, or 16, or 23 (assuming it attains these ages) is still 17. In a statistical sense, it is no closer to death after it has lived a long time than when it was born. Imagine a process the duration of which is determined by the rolling of a fair die. For the sake of concreteness, suppose the die is rolled every minute and the rule of existence of our process is that it will terminate on the first roll of a 3. Inasmuch as the probability of rolling a 3 on a fair die is 1/6, the expected "waiting time" for the occurrence of a 3, which is the reciprocal of the probability, is 6 rolls. So at the beginning of the life of our imaginary process, its life expectancy is 6 minutes. Now suppose it has already lived for 5 minutes, which is to say the die has been rolled 5 times and has not yet come up 3. If it is a fair die, the probability of it coming up 3 on the next roll is still 1/6 and consequently the life expectancy when the process is 5 minutes old is still 6 minutes—6 additional minutes not counting those already lived. So we can say that the process is no more likely to expire soon after having lived for 5 minutes than it was when it began. Obviously this argument generalizes, so we can say that no matter how long the process has lasted, its remaining life expectancy remains constant.

Although I have only observational evidence on the matter, I believe this to be a difficult idea for some people to accept. People who have trouble with the idea of a nonaging process are likely to point out that if a process terminates eventually it will do so at a particular point in time and it is nonsense to believe that one does not get closer to that point as time passes. In the case of our example, sooner or later the die will come up 3 and the process will terminate. Indeed if many such processes are started they will last for different periods of time, some for only a minute, a few for many minutes, but they will all terminate, and after about 6 minutes on the average. When one considers a specific process in retrospect, after it has terminated, one would not be inclined to say that after it had lived for *n* minutes it was no closer to its demise than when it began. A process that lasted for 9 minutes, for example, was 9 minutes from death when it began and only 4 minutes from death after it had existed for 5.

What has to be reconciled here are a probabilistic and a deterministic perspective. The deterministic perspective, which is easy to take in retrospect, locates events at specific points on a time line and therefore provides a basis for talking about distances between these events as though they were known or determinable. The probabilistic view focuses on what can be known with the information in hand. To say that a process is no closer to death at point B than at A, B being a later point in time, is to say that nothing has happened to increase the likelihood of its imminent demise, or to cause one's estimate of the additional time it is likely to live to decrease.

Which view is correct? I am not sure that one can claim that either of them is correct in any absolute sense. One can look at the situation either way. To the extent that one's behavior is to be influenced by the remaining life expectancy of a nonaging process, however, only the probabilistic view makes sense, inasmuch as our knowledge of the past in this case gives us no clue as to what will happen in the future. To put the matter in concrete terms, suppose that you are engaged in a gamble regarding when a nonaging process will terminate. If, over a series of bets, you consistently change your wager depending on how long the process has lasted, against an opponent who understands probability theory, and in particular, the concept of a nonaging process, you will lose.

The possible informativeness of the passage of time has been a complicating factor in the design of certain types of psychological experiments in which it is important that the participant's momentary expectation for the occurrence of a stimulus be held constant. Suppose one wishes to determine the effect of temporal uncertainty on reaction time to an auditory stimulus. In one condition—low temporal uncertainty—the signal is equally likely to occur at any time during a 2-second interval; in another condition—high uncertainty—it is equally likely to occur at any time during a 10-second interval. The problem is that, in both cases, as time passes during the interval, the probability that the signal will occur very soon increases; if, for example, 1.8 seconds of the 2-second interval have elapsed and the signal has not yet occurred, one can be sure it will occur during the next 0.2 second. The solution to this problem is to make the waiting time nonaging, which means to keep the instantaneous probability of signal occurrence constant over time. This can be done either by sampling waiting times from an exponential distribution, or by quantizing time and letting a computer decide with a constant probability, during each time quantum, whether to present a signal at the end of that quantum (Nickerson, 1967; Nickerson & Burnham, 1969). Temporal uncertainty can be manipulated in this case by varying what the instantaneous probability is; the important thing is not to let it change during a single experimental trial.

SUMMARY

Randomness is an elusive concept. Experts are not agreed on precisely what it means or how to determine whether something is random or not. The term can have different connotations as used by different people, or even when used by the same person in different contexts. It is not clear that people who use it always know what they mean by it in specific instances. I venture to guess that many people use the term freely without giving much thought to precisely how they would define it if asked.

Despite such problematic aspects of the concept, the idea of randomness is fundamental to the theory of probability. People who use it, as either writers or readers, should understand its complicated and somewhat controversial nature. They need to be aware of the different connotations it can be given and of the various tests that have been prescribed for its presence. Familiarity with the distinction between random processes and random products is especially important, and awareness of the some of the more common misunderstandings of randomness and closely related concepts should be helpful.

CHAPTER 3

Coincidences

Of all the abuses of mathematics, of all the abuses of science generally, no single phenomenon causes more misunderstanding than coincidences.

—Dewdney (1993, p. 40)

Strange coincidence, that every man whose skull has been opened had a brain!

—Wittgenstein (1953/1972, p. 28)

Coincidences can be fascinating, entertaining, intriguing—and sometimes very informative. We are fascinated when we bump into someone from our home town in a foreign city, when we learn that a new acquaintance has three children with the same first names as our own, or when we hear from an almost forgotten friend about whom we have just dreamed. Journalists sometimes entertain us with accounts of coincidental similarities between famous persons. Parallels in the lives of John F. Kennedy and Abraham Lincoln, their assassins and the details of their assassinations, have captured the attention and imagination of several writers (Lattimer, 1966, 1980; Russel, 1973; T. R. Turner, 1993; Wrone, 1981):

> Lincoln was elected in 1860, Kennedy in 1960. Both were deeply involved in the civil rights struggle. The names of each contain seven letters. The wife of each president lost a son when she was First Lady. Both Presidents were shot on a Friday. Both were shot in the head, from behind, and in the presence of their wives. Both presidential assassins were shot to death before they could be brought to

> trial. The names John Wilkes Booth and Lee Harvey Oswald each contain 15 letters. Lincoln and Kennedy were succeeded by Southerners named Johnson. Tennessee's Andrew Johnson, who followed Lincoln, was born in 1808; Texan Lyndon Johnson was born in 1908. ("Compendium of Curious," 1964, p. 19)

The writer of the *Time* magazine article from which this quotation was taken pointed out that, in addition to the accurate correspondences, such as those mentioned, many others have been produced by tweeking the facts just a bit.

To a list similar to the one that appeared in *Time,* Gardner (1967) added the following observations:

> Both the Federal Bureau of Investigation and the Secret Service, had they been skilled in the prophetic aspects of numerology, would have been more alert on the fatal day.
>
> The digits of 11/22 (November 22) add to 6, and *Friday* has six letters. Take the letters *FBI,* shift each forward six letters in the alphabet, and you get *LHO,* the initials of Lee Harvey Oswald. He was, of course, well known to the F.B.I. Moreover, *Oswald* has six letters. Oswald shot from the sixth floor of the building where he worked. Note also that the triple shift of *FBI* to *LHO* is expressed by the number 666, the infamous number of the Beast.
>
> The Secret Service, an arm of the Treasury Department, likewise should have been more alert. Two weeks before the assassination the Treasury Department released a new series of dollar bills, a sample of which I enclose. [This is all presented by Gardner as a letter he received from the famous numerologist, "Dr. Matrix."]
>
> Observe that this series is designated by the letter *K* on the left. In 1913, half a century earlier, when the Federal Reserve districts were designated, Dallas was assigned the letter K, the eleventh letter of the alphabet. For this reason "Dallas, Texas," where Kennedy was murdered appears beneath the "K." *Dallas, Texas* has eleven letters. *John Kennedy* has eleven letters.
>
> The serial number on this bill, as on all K bills, begins with K and ends with A—"*K*ennedy *A*ssassination." Beneath the serial number on the right is "Washington, D.C.," the origin of the Presidents's fatal trip.
>
> Below the serial number on the right, as well as above and below the serial number on the left, are two *11*'s. Eleven, of course, is the month of November. The two *11*'s add to 22, the day of the assassination. To the right of Washington's picture is "Series 1963A," the year of the assassination. (pp. 45–47)

It is amazing indeed, is it not, that the FBI and the Secret Service could have been so blind to the obvious pointers to the coming tragedy! Gardner's spoof is good fun, but the "coincidences" and the interpretations Dr. Matrix makes of them are not much more far-fetched than some that have been taken quite seriously in other contexts by people with a numerological bent.

WHAT IS A COINCIDENCE?

> "And we are both widows too!" said Barbara's mother. "We must have been made to know each other."
>
> "I haven't a doubt about it," returned Mrs. Nubbles. "And what a pity it is we didn't know each other sooner."
>
> "But then, you know, it's such a pleasure," said Barbara's mother, "to have it brought about by one's son and daughter, that it's fully made up for. Now, ain't it?"
>
> To this, Kit's mother yielded her full assent, and, tracing things back from effects to causes, they naturally reverted to their deceased husbands, respecting whose lives, deaths, and burials, they compared notes, and discovered sundry circumstances that tallied with wonderful exactness; such as Barbara's father having been exactly four years and ten months older than Kit's father, and one of them having died on a Wednesday and the other on a Thursday, and both of them having been of a very fine make and remarkably good-looking, with other extraordinary coincidences. (Dickens, 1840/1894, p 301)

My dictionary *(Webster's New Collegiate)* defines *coincidence* as "the occurrence of events that happen at the same time by accident but seem to have some connection." Diaconis and Mosteller (1989) give a very similar working definition: "A coincidence is a surprising concurrence of events, perceived as meaningfully related, with no apparent causal connection" (p. 853). Owens (1992) describes a coincidence as an event that has no cause: "A cause ensures that its effects are no coincidences—so whatever is a coincidence necessarily has no cause" (p. 2), or as "an event which cannot be explained" (p. 17).

For purposes of this discussion, I would qualify these conceptions slightly. First I want not to restrict the notion of co-occurrence or concurrence of events to events that happen at the same time. The fact that Joseph Frederick Johnson's full name is the same as that of his wife's maternal grandfather would probably be considered a coincidence, according to common usage of the term, although it does not involve the co-occurrence of events in a temporal sense. I will use the term co-occurrence in what follows, but will mean to include by it events that are seen as connected but do not necessarily occur at the same time. Also, the same illustration makes the point that *event* is to be given a fairly broad connotation.

Second, the notion of an event that has no cause needs sharpening. Leaving some of the strangeness of quantum mechanics aside, the prevailing view of how things work in this universe includes the idea that nothing that happens happens without a cause. Every event is embedded in a sequence of cause–effect relationships; it is a consequence of causes and becomes a cause in turn of subsequent events. Perhaps we might say that each of the events that comprise a coincidence has a cause, but that the cause of one of them is independent of that of the other. I run into a high school classmate in a restaurant in Tokyo; my being at that spot at that time has a cause, as does his being there at that time, but the causes are not related. We would say that for each of us there is a causal explanation as to why he is there, but neither of these explanations sheds any light on why the other person is there at the same time. It is in this sense that we can say that our meeting was a matter of chance.

WHAT MAKES COINCIDENCES INTERESTING?

What makes a coincidence interesting? And what *should* do so? We have to recognize first that people undoubtedly differ in the degree to which they attend to coincidences or find them fascinating. Paulos (1998) argues that differences in the way people react to coincidences—whether they are willing to accept most of them as insignificant or insist on always finding a meaning behind them—reveals something about their personalities and world outlooks. True or not, I suspect that most of us find at least some of the coincidences we witness to be sufficiently interesting to make us cast a thought or two in their direction.

Meaningfulness

All of us constantly witness very low probability co-occurrences, and pay them no mind; they do not even register as coincidences in the usual sense of the term. If someone points out to me that the co-occurrence of a blue jay landing on the bird feeder outside my window and the radio beginning to play *Claire de Lune* at the precise moment that I walked by the window on my way to get a cup of coffee was very unlikely a priori, I would have to admit that that was so, but normally I would not be surprised by such a co-occurrence—I would not even notice it as an interesting coincidence because the co-occurring events bear no meaningful relationship to each other in my mind. Our days are filled with such meaningless co-occurrences that go unnoticed.

Perhaps the most obvious difference between coincidences that capture our attention and those that do not is that in the former case the co-occurring events are meaningfully related in the observer's mind in some way. The appearance of a blue jay on the feeder outside one's window and the playing of *Claire de Lune* on one's radio are not, under most circumstances, likely to be meaningfully related events in most people's minds, so such a coincidence goes unnoticed. In contrast, the co-occurrence of the appearance of a blue jay on one's feeder and the playing of *Rhapsody in Blue,* while no less improbable than the coincidence of the blue jay and *Claire de Lune,* is much more likely to capture one's attention, because the two events in this case are more likely to be related in one's thinking.

Assumed Small Prior Probability

Presumably the smaller the retrospectively perceived a priori probability of the co-occurrence, the greater the surprise when it occurs. Expected co-occurrences are not surprising, nor are co-occurrences that appear to have been inevitable after the fact. Perceived probability is a function, in part, of the number of ostensibly independent events that coincide (Hakohen, Avinon, & Falk, 1981, reported in Falk, 1989): The discovery that two other people in a small group had the same birthday as oneself, for example, would be more surprising than the discovery that one did. As the definitions proposed by Diaconis and Mosteller (1989) and by Owens (1992) suggest, a coincidence is surprising, or continues to be surprising, only to the extent that one cannot find a causal explanation for it. One's initial surprise at unexpectedly running into several college classmates at a restaurant several decades after graduation would dissipate quickly upon learning that the happening was engineered by one of the people gathering.

But though small a priori probability of occurrence seems essential if a coincidence is to be interesting, it does not suffice to make it so. We witness coincidences that have infinitesimally small a priori probabilities every day and think nothing of them. My 2000 green Buick from Massachusetts arrives at a multibooth toll station on an interstate highway at the same time as do a red 1997 Chrysler from Montana, a silver 2002 Ford from Maine, a red 1992 Volvo from Connecticut, and a white 2001 Toyota from Florida, surely a low-probability confluence of events, but not one that anyone is likely to consider worthy of notice.

There are other explanations also of why coincidences surprise and interest us. Falk (1981–1982) suggests, for example, that when one experiences a coincidence, one is likely to focus on the specific happening as such and not to perceive it as one of a set of possible happenings any one of which would have been equally surprising if considered in isolation. When unexpectedly running into an old friend, for example, one is likely to focus on the low probability of a chance meeting of that particular friend at that particular spot at that particular time rather than on the fact that a chance meeting of some old friend at some place at some time during one's life is perhaps highly probable, and that *any* such meeting would be seen as unlikely when it occurs.

Personal Involvement

Falk (1989) and Falk and MacGregor (1983) obtained evidence that people find it somewhat easier, or more natural, to perceive coincidences in the latter way when they are experienced by other people than when experienced by themselves. People express more surprise, for example, when a coincidence involves them than when a comparable, or even identical, coincidence involves others. Falk was able to show that this is not simply a consequence of people selecting own coincidental events to report that were subjectively surprising to themselves but not to others. She found that even meaningless contrived coincidences were perceived as more surprising when they involved oneself than when they did not.

Given that co-occurrences that we perceive as interesting coincidences capture our attention when we experience them, it is not surprising to find that they tend to be remembered rather well (Hintzman, Asher, & Stern, 1978). Because they are remembered well, it may appear, in retrospect, that one has experienced more of them than one might have expected to by chance. This illustrates one way in which selectivity may play a role in making attention-getting chance coincidences seem to be more common than they are. Selectivity is also involved when we focus on low-probability co-occurrences that are of special interest to us and ignore those (much more frequent ones) that are not. And it may affect our impressions of the incidence of such events in general because only those co-occurrences that *are* of special interest to people who experience them get reported—the fact that person X found himself sitting next to person Y (whom he did not know and in whom he had no interest) in a theater in London on September 14, 1986, is unlikely to be considered noteworthy, despite the fact that the a priori probability of this co-occurrence must be considered extremely low.

THE COMMONNESS OF SURPRISING COINCIDENCES

> A tendency to drastically underestimate the frequency of coincidences is a prime characteristic of innumerates, who generally accord great significance to correspondences of all sorts while attributing too little significance to quite conclusive but less flashy statistical evidence. (Paulos, 1990, p. 35)

This quote from Paulos alludes to the fact that some coincidences may be much more highly likely than is generally realized, simply on the basis of chance. Consider, for example, the probability that within any modest-size group there are two people who have the same birthday. Imagine a gathering of four people. What is the probability that among these people there are at least two who celebrate their birthdays on the same day? Inasmuch as each person could have been born on any of 365 days, there are 365^4 possible arrangements of the four birthdates. There are $365!/(365-4)!$, or $365 \times 364 \times 363 \times 362$, ways in which the birthdates could be distributed so that *no two* are the same. So the probability that at least two of the four people have the same birthday is $1-[365!/(365-4)!]/365^4$. In general, the probability that among *n* people at least two have the same birthday is $1 - [365!/(365 - n)!]/365^n$. The answer for $n = 4$ is .016, not very likely. However, for moderately large n the probability is much larger; it is greater than .5 for n greater than 23.

Since it was first identified as an interesting phenomenon relating to probability theory, the chance coinciding of birthdays has received considerable attention in the statistical literature. Diaconis and Mosteller (1989), who review some of this literature, mention von Mises (1939) as perhaps the first to discuss the phenomenon. They point out too that the problem has been generalized in various ways and formulas have been developed to compute (or approximate) the probability that two or more of N people have the same birthday plus or minus one day, the probability of k or more of N people having the same birthday, and so on. B. Levin (1981) developed an algorithm for computing the smallest number of people, N, required to ensure a probability of at least 0.5 that at least k of them have the same birthday. Diaconis and Mosteller give $N \approx 47(k - 1.5)^{3/2}$ as a function that approximates the values produced by Levin's algorithm for k smaller than 20. According to Levin's computation, one needs a group of 1,181 for the probability that 10 or more people will have the same birthday to be greater than 0.5. For $k = 10$, Diaconis and Mosteller's approximation of N gives 1,165.

When should coincidences be surprising? What about the kinds of coincidences that were mentioned at the beginning of this section—learning that a new acquaintance has three children with the first same names as one's own, or hearing from an old friend about whom one has just dreamed? Each of these events seems highly unlikely, and therefore surprising when it occurs. But what we easily overlook when we think about such events is the very large number of things that could occur in one's life, any one of which might be considered a surprising coincidence. The a priori probability that a specific one of them will occur may be very small, but the probability that *some* one or more of them will occur is very high; or to make the point the other way around, the probability that *no* such events will occur is very low. When such an event does occur, we are more likely to be impressed by the low a priori probability of that particular event than by the high probability of the occurrence of *some* attention-getting low-probability event.

The point is that the fact that coincidences occur should not be surprising. To the contrary, we should be amazed if they did not occur. The example of the coinciding birthdays is but one of many that could be used to make the point. The example serves also to sharpen an important distinction. The probability that any two randomly selected people will have the same birthdays is very small, less than .003, but as we have seen, the probability that *some pair* of persons among a set of n people will have the same birthday is fairly large when n is on the order of a dozen or so.

Diaconis and Mosteller (1989) express the inevitability of low-probability consequences with "the law of truly large numbers," according to which "with a large enough sample, any outrageous thing is likely to happen" (p. 859). They point out, for example, that if a once-in-a-million coincidence occurs to one person in a million every day, on the order of 100,000 such occurrences a year would be expected in a population of 250 million people. (A similar argument was made by Alvarez [1965] in a letter to the editor of *Science* regarding parapsychology.) On the same logic, one-in-ten-thousand or one-in-one-thousand coincidences, any of which might seem remarkable from a narrow perspective, would be very common from a broad one. The same logic can bring us to the conclusion that any given individual is likely to experience some number of such low-probability coincidences over the course of a lifetime. In being surprised by the occurrence of specific low-probability events, one is making what Falk (1981–1982) calls the selection fallacy: One attends to a low-probability coincidence after the fact and is surprised by it only because one does not recognize that some such events are bound to occur.

In view of the high probability of low-probability co-occurrences, why are we surprised when they happen? The answer that comes most immediately to mind is that we are unaware, or forget, that the likelihood of such co-occurrences is high. As Gardner (1979) puts it, "Even mathematicians can forget that if enough people doodle long enough with random sequences of digits, it is highly probable that they will find highly improbable patterns. It is because most people fail to grasp this basic notion that they are unduly impressed when, out of the billions upon billions of possible ways coincidences can arise in daily life, one does occur" (p. 25).

Despite the fact that we should be surprised if coincidences did not occur, their occurrence can still be attention getting. A few days ago, I took a leisurely meandering ride on my bike on a particularly pleasant afternoon. I noticed at one point on the trip that my odometer was turning to 21 at precisely the moment that I was passing a roadside marker that read "mile 21." This was the only marker I saw on the trip. If I were more superstitious than I am, I might have read some significance into this coincidence. I am not that superstitious, quite, nevertheless, the event was sufficiently attention getting to have been retained as a memorable aspect of the trip.

CHANCE COINCIDENCES AND COMMON-CAUSE COINCIDENCES

Some coincidences seem best considered chance events; others are of interest because of the possibility that they are consequences of hidden relationships. The coincidence of a motive, an opportunity, and lack of an alibi, for example, can figure significantly in a criminal investigation. Such coincidences beg for an explanation. Believing that they happened by chance strains one's credulity.

The question of how to tell chance coincidences from coincidences that are evidence of a common cause is not easy to answer. Presumably it makes sense to look for a nonchance explanation of what appears to be a coincidence when it is reasonable to suspect that the coincidental events are causally related in some way. This criterion is quite subjective, and what one person may see as a clue to a causal connection another may be willing to write off as nothing more than a chance confluence of events.

Correspondences that can be perceived as chance coincidences are of little interest beyond the fact of their existence. Consider, for example the following facts. The Boer War ended in 1902 and the First World War began in 1914. Nineteen hundred and fourteen happens to be the sum of 1902 and the individual digits that comprise that number: 1902 + 1 + 9 + 0 + 2 = 1914. The First World War ended in 1919 and the Second World War started in 1939. Nineteen hundred and thirty-nine happens to be the sum of 1919 and the individual digits that comprise that number: 1919 + 1 + 9 + 1 + 9 = 1939. Having noted these coincidences, one has probably said all that needs to be said about them.

Or consider the fact that the 46th word of the 46th Psalm of the King James Bible is *shake,* while the 46th word from the end of the Psalm is *spear.* This conjunction might have gone unnoticed even by the ardent numerologists who discovered it if it were not for the fact that the King James translation of the Bible was completed in 1610, the 46th year of Shakespeare's life. Only someone who is committed to the idea that the Bible is full of coded messages and that every numerical relationship to be found in it has some significance, or someone who believes that King James' translators took the liberty of planting a cryptic tribute to the bard, is likely to make anything of it.

In *Gulliver's Travels,* Jonathan Swift has Laputan astronomers discover that Mars has two moons, one of which orbits the planet in the direction of the planet's rotation, making a complete revolution in about one third the time it takes the planet to rotate on its axis. These discoveries of Swift's fictional astronomers are remarkably close to the facts, but Swift wrote about 100 years before there existed a telescope large enough to see the moons of Mars. The chance correspondence between the rotation velocity of Mars and the orbiting speed of the satellite that orbits in the same direction is the more remarkable because this moon is the only known satellite in the solar system that revolves about a central body faster than the central body rotates. Velikovsky's explanation of this coincidence in *Worlds in Collision* is that Swift learned of the Martian moons from ancient manuscripts that recorded observations of the moons when Mars was close to the earth (Gardner, 1957). Gardner refers to the discoveries of Swift's Laputan astronomers as "perhaps the most astonishing scientific guess of all time" (p. 30).

There are many parallels between ancient Egypt and some of the precolonial cultures of Latin America, such as the knowledge of embalming and the building of pyramids. Some observers see such parallels as instances of discoveries or developments that occurred independently at about the same time in different parts of the world; others take them as suggestive evidence of communication between cultures or the effects of a common influence.

Often it is very difficult to determine whether a coincidence is due "merely" to chance or signifies some deeper relationship. And it can be hard to tell how much effort making this determination is worth in specific cases. It is not surprising therefore that particular coincidences are often interpreted differently by different observers. There are many examples in science of coincidences that have been sufficiently intriguing to some individuals to motivate years of research dedicated to understanding their significance, while being of little interest to, or dismissed as chance correspondences by, everyone else. Polkinghorne (1988) notes the importance of perspective here and suggests that what coincidences seem significant to us as individuals and motivate us to seek explanations is likely to depend upon "the interpretative scheme with which we approach the world" (p. 28).

COINCIDENCES IN SCIENCE

The noticing of coincidences has played an important role in scientific discovery and the development of scientific theory. The ability to see obscure connections—connections that most people miss—is believed by some to be among the more important abilities a scientist can have. Isaac Newton's postulation of the universal law of gravitation has been attributed to his notice of what was, at the time, a "very approximate, numerical coincidence," namely the fact that the parabolic path of a thrown rock and the path of the moon's orbit about the earth are particular cases of the same mathematical object (Wigner, 1960/1980). Apparent coincidences invite speculation in science as elsewhere, and finding explanations for coincidences has often been a goal of scientific research and theory building. The similarity of the eastern coastline of South America and the western coastline of Africa invites consideration of the possibility that the two continents were once joined. The fact that quarks and leptons are similar in certain respects, including the apparent organization of both in three groups, begs an explanation, and physicists are likely to see such a correspondence as the hint of a deeper causal connection. When the numbers 2.5029 and 4.669201 or their reciprocals keep popping up in studies of mathematical chaos, one is led to suspect that there is something special about these numbers in this context and that their frequent appearance is not "just" coincidental.

The distinction between coincidences in science that motivate search for a causal connection and those that do not can be subtle. The relative diameters of the sun and the moon and their relative distances from the earth (about 400-to-1 in both cases) are balanced in such a way that each body subtends about the same visual angle (1 degree) and therefore they have almost exactly the same apparent size to an observer on earth. (Although this coincidence has not motivated a search for a causal connection, it has proved to be useful to science by, for example, permitting the testing of a prediction of Einstein's theory of relativity during an eclipse of the sun.) The period of revolution of the sun on its axis and the period of revolution of the moon about the earth both are about 28 days. These are examples of coincidental relationships revealed by science to which scientists have attached no particular significance. They are, one might say, curiosities, but not of the type that motivate most people to seek causal explanations. In contrast, the fact that the orbits of the planets of our solar system are, with one exception, all in about the same plane is too great a coincidence for anyone to believe that there is not some causal connection and any theory of the origin of the solar system that is to get serious consideration among scientists must take this fact into account.

Commenting on the numerous regularities involving atomic weights and other characteristics of the elements that had been pointed out by his fellow chemists, Mendeleev quoted approvingly Strecker's (1859) observation that "it is hardly probable that all the above mentioned relations between the atomic weights (or equivalents) of chemically analogous elements are merely accidental" (Mendeleev, c. 1869/1956, p. 915). Mendeleev went on to produce the periodic table of elements for which we remember him today. The history of science provides many similar examples of "coincidences" that turned out, upon further investigation, to be anything but coincidental in the usual sense of the term.

A particularly interesting illustration of the fact that advances in scientific theory have sometimes provided new ways to view previously unexplained coincidences relates to inertia and gravity. It had been known at least since the days of Newton that the force of gravity on an object is proportional to the object's inertia. Lederman (1993) credits Newton with being the first to recognize that the *m* of

$$F = ma,$$

that represents the *inertial* mass, the stuff that resists force, and the *M* in

$$F = G(M_1M_2)/R^2,$$

that represents *gravitational* mass, the stuff that exerts pull on another object, are equal, or very nearly so. He refers to the equality of *m* and *M,* which are sufficiently different entities that they should not have the same name, as "an incredible coincidence" that "tormented scientists for centuries," and notes that Einstein incorporated it into his general theory of relativity (p. 96). Berlinski (2000) points out that the concepts of inertial mass and gravitational mass pertain to different parts of Newton's system: "Newton's second law of motion draws a connection between force, *inertial* mass and acceleration. There is no appeal to gravitation. Newtons's universal law of gravitation, on the other hand, draws its connection between force, *gravitational* mass and acceleration. There is no appeal to inertia. Two different conceptions of mass; two different laws of nature" (p. 201).

Recognition of the dual role of mass was articulated in the *principle of equivalence,* which refers to the equality of mass as a measure of resistance to acceleration and as a source of gravitational attraction. Newton himself saw the coincidence as remarkable and mysterious, and offered no explanation for it. Einstein resolved the mystery by declaring gravity and inertia to be the same thing and making this equivalence the center of the general theory of relativity, according to which there is only one type of mass and it is a consequence of the curvature of space.

These cases illustrate that attempting to account in a causal way for what might appear, at first, to be chance coincidences has been a fruitful endeavor in science. But one can find examples of attempting to account for coincidences that are generally considered not to be science at its best. A well-known case in point is that of the attention given, beginning in the mid-19th century, to the great pyramid of Egypt. Much of the fascination that modern observers have had for this structure is due to certain mathematical relationships noted first by John Taylor (1859) and shortly later by Charles Smyth (1865). Taylor and Smyth discovered numerous intriguing coincidences in measurements they made—for example, that the ratio of twice the pyramid's base to its height was roughly the same as the ratio of the diameter of a circle to its circumference; that the ratio of the pyramid's base to the width of a casing stone was 365, the number of days in a year; and that the pyramid's height multiplied by 10^9 was approximately equal to the distance from the earth to the sun. The discovery of these and many other correspondences provided von Däniken (1969) a basis for arguing that the earth had been visited by intelligent extraterrestrials in the past.

What is the critical difference between this work with pyramid measurements and the events mentioned in connection with Newton, Mendeleev, and Einstein? In each of the latter cases, a coincidence was noticed that seemed to call for a search for a common physical explanation, and the explanations that were proposed were subject to test by virtue of their predictive implications regarding observable phenomena. In the case of the pyramid measurements, many relationships were considered and only those that were deemed to reflect interesting correspondences were given further attention. This kind of selectivity lends itself to underestimation of the probability that the coincidences noted could be due to chance.

Gardner (1957), who describes this work, refers to Smyth's book as a classic of its kind illustrating beautifully "the ease with which an intelligent man, passionately convinced of a theory, can manipulate his subject matter in such a way as to make it conform to precisely held opinions" (p. 176). As Gardner points out, a complicated structure like the pyramid provides a great many opportunities for measurements of length and those measurements and the results of calculations based on them are bound to coincide here and there with numbers that are of interest for historical or scientific reasons, simply because the set of numbers to be considered is so large. If one has the freedom to pay attention only to those measurements or calculations that yield something of interest and ignore the great many that do not, one is pretty sure of finding some correspondences by chance: "Since you are bound by no rules, it would be odd indeed if this search for Pyramid 'Truths' failed to meet with considerable success" (p. 177).

Kepler's Spheres

One of the better known examples of a coincidence that fired the imagination of a great scientist is Kepler's discovery of a correspondence between the orbits of the planets of the solar system and the regular polyhedra. Since before Plato it had been known that there are five such shapes: the tetrahedron (pyramid; 4 triangular faces), the hexahedron (cube; 6 square faces), the octahedron (8 triangular faces), the dodecahedron (12 pentagonal faces), and the icosahedron (20 triangular faces). In Kepler's day there were only five known planets in addition to the earth. Kepler was convinced that these "perfect" forms held a key to a fundamental aspect of the solar system's design. He labored to find the connection between the polyhedra and the planetary orbits, and eventually he believed he had succeeded in doing so.

For each regular polyhedron, imagine two spheres, the largest sphere that can be enclosed within the polyhedron and the smallest sphere in which the polyhedron can be contained. For any given polyhedron the ratio of the diameters of the inner and outer sphere is a constant; that is, it is independent of the size of the polyhedron. This ratio differs however for the different polyhedra. Kepler discovered that if the polyhedra are scaled in size and nested in a specific order in such a way that the inner sphere of one polyhedron coincides exactly with the outer sphere of the next smaller one in the nesting, the ratios of the diameters of the nested spheres approximate the ratios of the diameters of the planetary orbits.

Here is his own account of the connection he worked out and his elation upon finding it, as quoted in Boorstin (1985, p. 310):

> The earth's orbit is the measure of all things; circumscribe around it a dodecahedron, and the circle containing this will be Mars; circumscribe around Mars a tetrahedron, and the circle containing this will be Jupiter; circumscribe around Jupiter a cube, and the circle containing this will be Saturn. Now inscribe within the earth an icosahedron, and the circle contained in it will be Venus; inscribe within Venus an octahedron, and the circle contained within it will be Mercury. You now have the reason for the number of planets.... This was the occasion and success of my labors. And how intense was my pleasure from this discovery can never be expressed in words. I no longer regretted the time wasted. Day and night I was consumed by the computing, to see whether this idea would agree with the Copernican orbits, or if my joy would be carried away by the wind. Within a few days everything worked, and I watched as one body after another fit precisely into its place among the planets.

Kepler considered the fact that there are only five regular polyhedra to be the *reason* why there were only six planets. He had, Butler (1970) suggests, "thought that since there could only be five regular solids, thus disposed, he had shown conclusively why there were six and only six planets, echoing Bongo's view that numerological arguments reveal to us the 'why' of things" (p. 87). The Bongo to whom Butler refers is Pietro Bongo, who wrote during the 16th century on number symbolism, or number mysticism.

How surprised should we be by the correspondence that Kepler discovered? Is this one of those coincidences that seems to demand an explanation in terms of a common cause? We need to distinguish here between surprise at the fact that Kepler thought there might be some connection between perfect solids and planetary orbits and surprise at the "closeness of fit" that he found. Let us consider first the question of how surprised we should be that Kepler was looking for a connection between the solids and the orbits.

The regular polyhedra—variously known as the "Platonic solids," the "perfect solids," or the "cosmic bodies"—had been objects of great fascination among philosophers since their discovery at least 2,000 years before Kepler and they had figured prominently in explanations of natural phenomena among the classical Greeks. The Pythagoreans, who thought the number of such figures was four (the dodecahedron had not yet been discovered), associated them with the four primal elements—earth (hexahedron), air (octahedron), fire (tetrahedron), and water (icosahedron). By the time of Plato the dodecahedron had been discovered and it represented to him the universe as a whole. It really is not surprising that Kepler, with his mystical perspective, his propensity to see beauty in regularity, and his firm conviction that the universe and all it contained were the perfect work of an intelligent architect, should think to look for a correspondence between the five perfect mathematical objects that he knew and the orbits of the planets that, to him, comprised the entire solar system.

Whether we should be surprised at the closeness of the correspondence (not exact) between the ratios of Kepler's nested spheres and the ratios of the planetary orbits is a different matter altogether. This is a question of statistical reasoning. How should we think about this question? The basic data of interest are two sets of ordered ratios. One of these sets of ratios is derived from the diameters of the planetary orbits, ordered from, say, largest to smallest. The first ratio in the set is the ratio of the diameter of the largest orbit to that of the second-largest; the second one is the ratio of the diameter of the second-largest orbit to that of the third-largest; and so on. The second set is derived in an analogous way from the diameters of the nested spheres containing the perfect solids. The first ratio is the ratio of the diameter of the largest sphere to that of the next-largest sphere, and so on. But which polyhedron should determine the largest sphere?

Here is a very important point. Five polyhedra can be ordered in 120 different ways, so Kepler had a choice of 120 different nesting arrangements, each of which would produce a unique set of ordered ratios against which to compare the ratios derived from the planetary orbits. Did Kepler pick the one that produced the best fit? I presume that he did. We would be more impressed by the fit that Kepler found if it had been produced by arranging the polyhedra in a natural order, say from most simple to most complex or vice versa. There seems to be no rationale for the order that Kepler selected, except, presumably, the fact that that order came closest to yielding the correspondence for which Kepler was looking. Today this would not generally be considered acceptable form, but probability theory and statistics were not established fields of mathematics at the time Kepler did his work. The book in which Kepler advanced his ideas about the correspondence between the planetary orbits and the perfect polyhedra was published in 1596, 61 years before Huygen's publication of the initial work of Pascal and Fermat on probability and 67 years before the appearance of Cardano's *The Book of Games of Chance*. And, as Devlin (2000a) points out, "in seeking to understand the patterns of nature through the abstract patterns of mathematics [he was] working within a tradition that continues to this day to be highly productive" (p. 152).

Coincidences Between Mathematics and Physics

According to a well-known theorem in number theory,

$$\sum_{i=1}^{n}(2i-1)=n^2,$$

The equation expresses the fact that the sum of the first n odd positive integers is equal to n squared. I mention this here in order to point out the correspondence between this relationship and Galileo's discovery of the law that relates the distance that a falling body travels to the duration of its fall.

An object falling under the force of gravity increases its speed of travel as it falls. If we represent how far it falls during the first second after it is dropped as 1 dfs (distance in first second) and measure the distance traveled in each successive second thereafter, we will find that it falls 1 dfs during the first second, 3 during the second, 5 during the third, and $2n - 1$ during the nth. Inasmuch as the sum of the first n odd integers equals n^2, we can represent the relationship between distance traveled and time as $d = t^2$, where t is in seconds and d in units of dfs. Usually the relationship between distance traveled and time is written as $d = kt^2$, the k being necessary because the unit of distance used is not the distance traveled during the first unit of time. Galileo discovered this relationship in the 16th century and demonstrated it in an ingenious way by rolling balls down an inclined plane.

The point of interest in the present context is not the fact that Galileo made this discovery, which is not to deny that it was a great discovery, but the fact that the distance traveled by a falling object and the time it has been falling is such an elegantly simple one. What do squares and falling objects have in common that they should be describable, not approximately but precisely, in the same terms? If we did not already know, thanks to Galileo, the relationship between distance and time and were about to do his experiment for the first time, would there be any reason to expect nature to behave in such a meticulously neat way? Might we not be inclined to expect that the distance traveled during the second second would be some noninteger multiple—say, 1.5314 ..., or 2.9273 ..., or even 3.0001 ...—of that traveled during the first? But 3 exactly? More generally, is there anything that would lead us to expect that the distance traveled in t seconds would turn out to be precisely proportional to t^2?

In fact, we could convince ourselves, without making any measurements, that distance traveled must increase with the square of time *if* we started with the assumption that a falling object picks up speed at a uniform rate, which is to say that its acceleration is constant. On this assumption, the average speed of the object during any interval of time is the average of its speed at the beginning of the interval and its speed at the end of it. Given that the distance traveled during any interval is simply the average speed during that interval times the duration of the interval, the assumption of a constant rate of change of speed allows us to express the distance traveled during an interval as a simple function of the speeds of the object at the beginning and end of that interval. Thus the distance traveled during an interval of duration t, given a speed of s_b at the beginning of the interval and one of s_e at the end would be $(s_b + s_e)t/2$.

Suppose we arbitrarily divide time, conceptually, into successive intervals such that the speed of the falling object will increase by 2 feet per second (fps) during each interval. On our assumption of a constant rate of increase in speed, these intervals would all be of equal duration. Thus over the course of the first interval the speed, in feet per second, would go from 0 to 2, during the second it would increase from 2 to 4, during the third from 4 to 6, and so on. The average speed would be $(0+2)/2=1$ fps during the first interval, $(2+4)/2=3$ during the second, $(4+6)/2=5$ during the third, and $2n-1$ during the nth. Inasmuch as average speed is expressed in feet per second and duration is in seconds, seconds disappear from the computation of the distance traveled during any given interval, and that distance, in feet, is simply the average speed during that interval. So the object travels 1 foot during the first interval, 3 feet during the second, 5 during the third, and $2n-1$ during the nth. The *total* distance traveled by a falling object, from the moment it began its descent, would of course be the sum of the distances traveled during the successive intervals, so after the first interval it would have gone 1 foot, after two intervals $1+3=4$ feet, after 3, $1+3+5=9$ feet, and, given that we know that the sum of the first n odd integers is n^2, we can see that the distance traveled during n intervals will be n^2 feet.

So we have an explanation, of sorts, of the neatness of the relationship between the distance traveled by a falling object and time, and, in particular, of the fact that distance traveled increases with the square of time. It is an explanation at least in the sense that the relationship is somewhat less mysterious when looked at this way, and does not even require empirical measurement for verification. Given the assumption that speed changes at a uniform rate, it is clear that the relationship must be of this form. But what about the assumption that speed changes at a uniform rate? What gives us the right to make it? Why should we assume, before we look, that nature will behave in this way? One might say that changing at a constant rate is simpler than changing at a variable rate, but it is not clear what that might mean, except that we find constant change easier to think about and to describe than variable change. Surely nature does not select its laws for our conceptual convenience. On the other hand, no one yet has given a much better explanation of why mathematics—simple mathematics—is as descriptive of the world as it is, or at least appears to be.

Let us consider another use of mathematics to describe a physical phenomenon. Imagine a particle moving under gravity down a frictionless wire from point A to point B, where point B is displaced laterally some distance from point A. Determining what shape the wire must be in order to minimize the time of travel is an old problem in mathematics and one that was considered not only by Galileo (who thought the answer was an arc of a circle) but by such other luminaries as Jacob and Johann Bernoulli, Newton, and Leibniz. The solution is an inverted cycloid, which is the path traced by a point on the rim of a moving wheel. This seems at least as curious a coincidence as the one that has squares and gravity in the same box. What in the world does the minimum-time problem have to do with a point on the rim of a moving wheel? Why should both yield precisely, again not approximately, the same curve? It turns out that the same curve describes also the only path with the property that a particle moving along it under gravity will take the same time to reach a given point independently of its starting point. The latter fact was discovered by Christian Huygens a few years before the minimum-time problem had been solved. Is it hard to imagine how Johann Bernoulli could speak of being "petrified with astonishment" upon learning that the cycloid was the solution to both problems?

These examples are illustrative of countless coincidences that we find between mathematics and the physical world. Just to mention two others: We find correspondences between the properties of triangles and the properties of (sound-pressure, electromagnetic) wave phenomena and between the equations that describe the curves obtained by slicing a cone at different angles (the conic sections) and the trajectories of missiles on earth and of the heavenly bodies in space. I suspect that we do not wonder much about such coincidences, that indeed it does not occur to us that there is here anything to wonder about. But what, after all, do triangles have to do with sound-pressure waves? Or conic sections with bodies moving through space? Why should an area of mathematics that was developed to deal with three-sided figures or cones prove to be so useful in describing the behavior of vibrating air molecules or the meanderings of planets, comets, and thrown rocks? Everywhere one looks in mathematics and science one finds similar, and similarly surprising, correspondences.

Perhaps we should assume that the mathematics that we use in our everyday lives has been shaped by our thinking in such a way as to ensure conformity with our perception of the environment. But as Polkinghorne (1988) notes, this assumption would not account for the surprising applicability of abstract mathematics to newly discovered aspects of the world. It "does not begin to explain why highly abstract concepts of pure mathematics should fit perfectly with the patterns of the subatomic world of quantum theory or the cosmic world of relativity, both of which are regimes whose understanding is of no practical consequence whatsoever for humankind's ability to have held its own in the evolutionary struggle" (p. 21).

One of the intriguing questions prompted by the fact that physical reality is describable by simple mathematics is the question of whether reality is the way it is because that is the only way it can be. Do falling objects cover a distance that is proportional to the square of the duration of their fall because gravity *must* work that way—because any other relationship would be physically impossible? If that is the case, what dictates this necessity? And if it is not the case, why are such surprisingly simple relationships so common? The fact that the gravitational force between two objects varies precisely with the inverse of the square of the distance between them may be seen as evidence of the simplicity of natural law; the fact that the qualitatively different force that holds an electron to an atomic nucleus—the Coulomb force—has the same inverse square relationship is remarkable. Surely it is too much to believe that such commonalities are coincidences that have no rational explanation.

Large-Number Coincidences

> Accidental numerical relations between quantities as unconnected as the fine structure constants for gravity and electromagnetism, or between the strengths of nuclear forces and the thermodynamic conditions of the primeval universe, suggest that many of the familiar systems that populate the universe are the result of exceedingly improbable coincidences.... Turning to the subject of cosmology—the study of the overall structure and evolution of the universe—we encounter further cosmic cooperation of such a wildly improbable nature, it becomes hard to resist the impression that some basic principle is at work. (Davies, 1982, p. 75)

Among the coincidences that have increasingly attracted the attention of physicists, astronomers, and cosmologists in recent years are several that involve large, in some cases very large, numbers. Some of these large-number coincidences, though interesting, have prompted little comment; others have caused a great deal of speculation and debate.

An example of the first case is the coincidence between the gravitational energy in the universe and the universe's total mass: "When one adds up all the mass in the observable portion of the universe, a very large number is obtained. When one calculates the amount of negative gravitational energy that exists in the same region of space, another large number is obtained. As far as we can tell, the two numbers are about equal" (Morris, 1987, p. 141). Morris notes, however, that although this is an intriguing coincidence, no conclusions have been drawn from it.

A large-number coincidence that has puzzled physicists for many years involves what would appear to be two completely independent ratios that turn out to be almost identical. The ratio of the linear extent of the known universe to the diameter of an atom is roughly the same as the ratio of the electromagnetic force between an electron and a proton and the gravitational force between them. In both cases the ratio is roughly 10^{40}. Although associated with the names of Eddington and Dirac, this coincidence was apparently first pointed out in 1919 by Hermann Weyl (J. A. Wheeler, 1986). (The first ratio is sometimes expressed in terms of the Hubble time—about 10^{10} years—and the "nuclear time," which is the time that light takes to travel a distance equal to the diameter of a proton. Davies [1982] gives 10^{41} as this ratio. If one uses the linear extent of the universe and the diameter of a proton for one ratio and the gravitational attraction and electromagnetic repulsion between two electrons for the other, one gets ratios closer to 10^{42} [Feynman, 1965/1989].)

These two ratios are sometimes referred to as the first and second cosmic numbers, respectively. The fact that they are nearly equal is a surprising coincidence, inasmuch as there is no reason to expect them to be related in any way. And there is more. Whereas the second number is invariant, the first one changes as the universe expands, so the mystery of the coincidence is deepened by the fact that the numbers coincide at precisely the time that we happen to be around to notice the fact. As if this were not mystery enough, Dirac pointed out in the 1930s that the estimated number of particles in the observable universe is approximately the square of the coincidental ratio. Coincidences involving magnitudes of about 10^{40} are intriguingly common and have attracted the attention of both physicists and cosmologists. Discussions of several of them can be found in Davies (1982) and in Barrow and Tipler (1988).

Other Coincidences

There are numerous other coincidences the significance of which has been the subject of speculation and debate. An example is the coincidence between the laws that relate strength to distance for both electromagnetic and gravitational force. Although, as already noted, the electromagnetic repulsion between two electrons is greater by at least 40 orders of magnitude than the gravitational attraction between them, these forces both vary inversely with the square of the distance between the objects involved. To date, no one knows whether this coincidence is significant, but the joining of the forces of electromagnetism and gravity within a single theoretical framework remains a major goal of physics.

The earth has many properties that contribute to its habitability by life as we know it. These include its distance from the sun, the near circularity of its orbit, the tilt of its rotational axis, the size and position of its moon, its tropospheric ozone shield against ultraviolet radiation, its magnetic shield against potentially lethal cosmic radiation, the abundance of oxygen in its atmosphere and liquid water on its surface; the list is easily extended. Some people find it hard to accept this constellation of unusual properties that makes the earth conducive to life as purely coincidental and take it as evidence that the earth is, by design, a special place.

Taking the existence of the earth with all its properties as a given, there are other coincidences that are fortunate from the point of view of its inhabitants. A case in point is the fact that a human being happens to be about the right size to have developed the use of fire for practical ends. A fire that is smaller than a small campfire is difficult to keep going because it is so easily extinguished, and a much larger one is hard to keep under control: "Prometheus was just large enough to feed the flame and keep from getting burnt" (Stevens, 1974, p. 25).

Another example is the coincidence involving the freezing temperature of water, the fact that much of the earth has temperatures below this point for significant portions of the year, and the curious fact that, unlike most liquids, which contract upon freezing, water expands—not much, but enough to make ice float. This is fortunate because, if ice sank, lakes would freeze solid in winter—as the ice formed, it would sink to the bottom until the entire lake was a solid block—and many would thaw only close to the top in the summer. This would mean that much of the earth that is now habitable would not be, at least by life forms with which we are familiar that depend on an abundance of liquid water as a habitat.

When one considers the universe more generally, one also finds numerous quite remarkable coincidences, some of which appear to be essential to the existence of life, and others of which seem to be required for the existence of anything. These include the fact that gravity is at the right strength to permit the formation of planets (Carter, 1974); several remarkable properties of water in addition to its expansion upon freezing (Greenstein, 1988); the coupling of the resonance between helium and beryllium, which enhances the production of beryllium in red giant stars, and that between the reacting beryllium and helium nuclei with the carbon atoms that they form (if it were not for this unique double matching of resonances there would be no carbon production and consequently no carbon-based life; Barrow [1991] refers to this coincidence as "something akin to the astronomical equivalent to a hole-in-one" [p. 95]); the fact that neutrons outweigh protons by just enough to permit the existence of hydrogen and hence the existence of hydrogen stars, which, unlike helium stars, last long enough for life to develop (Davies, 1982); the precise correspondence between the expansive force of the big bang and the force of gravity (to within 1 part in 10^{60}) at the "Plank time" of 10^{-43} seconds after the big bang (Gribbin & Rees, 1989); the equivalence of the electrical charges of protons and electrons, which differ dramatically in almost every other respect (if these charges were not equal to an accuracy of one part in 100 billion, objects the size of human beings or smaller could not exist; an even greater accuracy is required to permit the existence of much larger structures, such as the earth or the sun) (Greenstein, 1988); the precise strengths of the strong and weak forces (small differences in either would rule out the possibility of the universe as we know it) (Barrow & Silk, 1983; Davies, 1982; Gribbin & Rees, 1989). And so on.

I mention these "coincidences" to make the point of their existence, but I will make no attempt here at an extensive discussion of what, to me, is a most intriguing topic. Thoughtful and thought-provoking discussions of these and other coincidences that are essential to life as we know it, and in some cases to the existence of a habitable universe or even any universe at all, can be found in Roxburgh (1977), Davies (1982, 1983, 1988, 1992), Hoyle (1983), Polkinghorne (1986, 1988), Barrow and Tipler (1988), Greenstein (1988), Gribbin and Rees (1989), and Barrow (1990, 1991, 1992), among other places. Is the identification of such coincidences simply another illustration of selectivity of the sort described earlier in connection with studies of the great pyramid, and are the coincidences that are identified to be expected on the basis of chance if properly viewed, or are they aspects of the universe that beg causal explanation. I confess to leaning strongly to the latter view.

There can be no question about the fact that chance coincidences can be made to appear to be more significant than they are by selectively focusing on them after the fact and failing to take account of the high probability that some such co-occurrences are bound to happen by chance. One might push this observation to the extreme of dismissing *all* coincidences that do not have obvious explanations as due to chance. This does not seem reasonable to me; in my view, there is a point at which attributing a coincidence to chance taxes credulity to a greater degree than assuming that there is a causal explanation to be found. Where that point is depends in part, I suspect, on one's philosophical and/or religious perspective as much as on anything else.

THE ANTHROPIC PRINCIPLE

> We are one species among millions on an undistinguished planet circling an undistinguished star that travels along an undistinguished orbit in an undistinguished galaxy. But we must acknowledge that our place in the picture, however small, is not insignificant, because it is we who make and remake the picture. And the strongest and most persistent reason for doing so—for trying to understand what the universe is really like—has always been the desire to understand how we fit into it. (Layzer, 1990, p. 231)

> The anthropic principle begins to look like a name for the repository in which we set aside all the things that physics cannot yet explain.... The anthropic principle is used to explain those things for which physics alone, we suspect, cannot provide an answer. It performs the role that less artful scientists in earlier ages ascribed unabashedly to a prime mover, or to God. (Lindley, 1993, p. 250)

> As we look out into the Universe and identify the many accidents of physics and astronomy that have worked together to our benefit, it almost seems as if the Universe must in some sense have known that we were coming. (Dyson, 1971, p. 59)

Davies (1982) refers to the discovery that many of the coincidences mentioned in the preceding section are necessary for our existence as one of the most fascinating discoveries of modern science. Certainly it has prompted a great deal of thought and discussion among scientists and philosophers in recent years. One outcome of this thinking that has received considerable attention is the *anthropic principle* or, as it is sometimes called, the *anthropic cosmological principle.*

The term *anthropic principle* was used first by the British physicist Brandon Carter in 1974. The basic idea has been elaborated by numerous others, notably G. J. Withrow and Robert Dicke. According to this principle, the very existence of human life constrains the universe. That is to say, given the existence of observers, the universe must have certain properties that are prerequisite to their existence. One statement of the principle is, "If some feature of the natural world is required for our existence, then [inasmuch as we exist] it must indeed be the case" (Greenstein, 1988, p. 46). Another similar expression of it is, "The only things that can be known are those compatible with the existence of knowers" (Greenstein, 1988, p. 47).

These assertions seem obvious, unlikely to provoke controversy, and not even very interesting. To be sure, if the conditions that permit life and knowers to exist did not hold, we would not be here to realize it; they must hold, given that we are here. But does this observation constitute an explanation of anything? Does it make the fact that the critical conditions do hold any less mysterious? For some at least, it does not. As Gribbin and Rees (1989) put it, "To say that we would not be here if things were otherwise ... need not quench our curiosity and surprise at finding that the world is as it is" (p. 285). Or as Davies (1992) argues, "The fact is we *are* here, and here by grace of some pretty felicitous arrangements. Our existence cannot of itself explain these arrangements" (p. 204).

A universe that is compatible with the existence of knowers appears to be extremely unlikely on a priori grounds, depending as it does on an impressive number of apparently independent, but precise conditions. Our surprise comes from the narrowness of the tolerances of the critical conditions, from the fact, in other words, that if this or that constant differed from its actual value by the tiniest amount, no universe containing knowers—indeed perhaps no universe at all—could exist. The existence of the universe and our existence in it would impress us less if they were known to be compatible with a wide range of conditions. Why, we feel compelled to ask, did precisely the right conditions pertain, when the probability of them doing so by chance seems to have been so infinitesimally small?

Weak and Strong Forms of the Principle

The anthropic principle has been stated in a variety of ways and means different things to different people. A distinction is often made between weak and strong versions of the principle. According to the weak anthropic principle (which is the interpretation given in the preceding paragraphs), life can exist only in an environment that is habitable. From the fact of our existence, therefore, we can infer certain properties of the universe, namely those properties it must have in order to permit our existence. For example, from the fact that we are a carbon-based form of life, we can infer that the universe must be, say, between 10 and 20 billion years old.

The argument involves assumptions about the origins and life cycles of stars, the synthesis of complex elements, the formation of the solar system, the evolution of life, and the times required for these processes, but the reasoning is relatively straightforward and easy to follow. If the universe were much younger, according to this argument, we would not be here, because there would not have been enough time for carbon and the other heavy elements that are necessary for human life to have been produced by the thermonuclear activity within the stars. When it is much older, the stars that are necessary to sustain life will have burned themselves out. We live during the only time relative to the universe's evolution, so this argument goes, that is conducive to life. A similar line of reasoning can be used to infer other properties that the universe must have.

The strong form of the anthropic principle holds that a habitable environment *must* exist; in Carter's words, "The universe must be such as to admit the creation of observers within it at some stage" (quoted in Davies, 1982, p 120). Davies points out that the strong anthropic principle, at least as articulated by Carter, has a quite different philosophical basis than does the weak one:

> Indeed, it represents a radical departure from the conventional concept of scientific explanation. In essence, it claims that the universe is tailor-made for habitation, and that both the laws of physics and the initial conditions obligingly arrange themselves in such a way that living organisms are subsequently assured of existence. In this respect the strong anthropic principle is akin to the traditional religious explanation of the world: that God made the world for mankind to inhabit. (p. 120)

Using the theory of quantum mechanics as a point of departure, some scientists have come to the position that the very existence of the cosmos and the matter that comprises it depend on the act of observation and therefore the existence of an observer. This is a strong form of the anthropic principle indeed. Greenstein (1988) has expressed this form of the principle in emphatic terms: "Why did the cosmos bring forth life? It had to. It had to in order to exist" (p. 198). "Apart from observation ... the electron has no objective reality at all. It has merely what might be called a set of potentialities, any one of which can be called into being. It is the observation itself that brings the physical world into existence" (p. 222). "Nothing exists unless it is observed" (p. 237).

Davies (1982) rejects the idea, which some interpretations of the strong principle seem to entail, that observers are responsible for the creation of the universe they observe: "Special physical conditions may produce man, but man can hardly be attributed the credit for establishing his own environmental requirements" (p. 122). But he does see the possibility of some justification for the strong principle in quantum mechanics: "Although the quantum observer cannot be said to actually create his own universe in the conventional sense of the word 'create,' an analysis of quantum measurement theory does open the door to providing a plausible physical, as opposed to philosophical, justification of the strong anthropic principle" (p. 122).

A problem with this form of the principle is that it seems to imply an infinite regress. According to it, existence depends on the act of observation. But the act of observation requires the existence of an observer. And the existence of this observer depends, according to the principle, on an act of observation, which requires another observer, and so on ad infinitum.

Is the Anthropic Principle Scientific?

Is the anthropic principle a scientific theory? Certainly it is a hypothesis about the nature of reality that has been proposed by scientists. And the reason it was advanced was to explain the numerous remarkable coincidences on which the existence of the universe and ourselves appears to depend. It fails, however, to meet the cardinal criterion of testability. It appears not to be falsifiable, in principle. As Davies (1982) notes, it is hard to see how it can be used to make a testable prediction, "because any physical theory that is inconsistent with our existence is manifestly incorrect anyway" (p. 129). Being compatible with any conceivable experimental outcome puts it in the same category as all other theories that explain everything and consequently explain nothing.

Pagels (1991) argues further that the anthropic principle has no influence on the development of contemporary cosmological models. No knowledge, he claims, has been gained as a result of its adoption, and most physicists and astrophysicists simply ignore it in the pursuit of their research: "I would opt for rejecting the anthropic principle as needless clutter in the conceptual repertoire of science" (p. 359). More emphatic dismissals of the principle can be found: Gell-Mann (1994), for example, says, "That idea seems to me so ridiculous as to merit no further discussion" (p. 212). There are, of course, many evidences in the scientific literature that an idea that appears absurd to one knowledgeable observer can be quite compelling to another.

What is one to make of all this? It will not do to say that one should take the scientific view, because scientists are not of one mind on the issue. We should note too that the motivation for the introduction of the anthropic principle was to account for coincidences that are, at least from some perspectives, strongly suggestive of design in the universe. In the words of the astronomer, Fred Hoyle, it all looks very much like a "put-up job." Commenting on the dependence of the formation of carbon on closely corresponding resonances, Hoyle argued that a "commonsense interpretation of the facts suggests that a superintellect has monkeyed with physics, as well as chemistry and biology, and that there are no blind forces worth speaking about in nature" (quoted in Davies, 1982, p. 118). Some scientists dismiss the idea of design in the universe out of hand. Some, as evidenced by this quote and others that could be cited, do not.

All we know about—all we *can* know about—is the universe in which we live. Whether that is the only universe that exists, that ever has existed, that ever will exist, we have no way of telling. The evidence that is available to us, or rather our interpretation of that evidence, tells us that our universe is on the order of 12 billion years old, give or take a billion or two, and finite in extent. One can imagine ours to be one of an infinity of universes, the vast majority of which came into existence under conditions other than the exquisitely balanced ones that permit the development of life and knowers. On this view, the fact that the parameters of our universe appear to be tuned precisely so as to accommodate our existence is not particularly interesting; one would expect such a combination of conditions to occur by chance many times—indeed an infinite number of times—given an infinitely large sample, and the universe in which we appear *must* be one of those for which they hold. One can also imagine the universe in which we live to be all there is or ever was. On this view, the fine-tuning of the parameters becomes very interesting indeed.

What one makes of the coincidences that are usually associated with the anthropic principle is likely to depend strongly on one's general perspective on the world and the presuppositions on which that perspective rests. An individual who believes that the world was brought into existence by an intelligent Creator will see the coincidences that have been the focus of the preceding paragraphs as the consequences of careful design. From this perspective, it is not surprising, for example, that the basic constants need to be precisely what they are in order to produce life: "It is just what one would expect if an immensely wise God wished to produce a life-bearing universe, if the whole thing was purposive" (K. Ward, 1996, p. 52). Ward argues that "it is not at all what one would expect, if it was a matter of chance. Every new scientific demonstration of the precision of the mathematical structure needed to produce conscious life is evidence of design. Just to go on saying, 'But it could all be chance' is to refuse to be swayed by evidence" (p. 52).

But one who believes that chance rules all may have no difficulty in attributing the coincidences to chance. One is unlikely to get much help from science in deciding this issue, inasmuch as it is not, in principle, decidable by scientific means. This is not to say that the choice of what to believe is a toss-up. One must interpret and weigh the evidences as best one can. And when one does so, it becomes, as Davies (1983) puts it, a matter of what one finds it easier to believe. I confess to finding it easier to believe in a cosmic Designer than in the idea that it all came about by chance. Indeed, precisely what having everything ruled by chance might mean is not entirely clear to me. Chance, if it is to be used as an explanatory construct, must be given a much more "deterministic" connotation than is generally intended when the term is employed.

SUMMARY

Coincidences are ubiquitous. They capture our attention when they unexpectedly bring together presumably independent events that we find interesting or meaningful in combination. Some coincidences are best considered chance co-occurrences; others can be traced to common, or closely related, causes. When noticed, coincidences often motivate search for causal explanation. Efforts to find explanations can evoke or reinforce superstitious thinking; they have also sometimes led to scientific discoveries. Certain remarkable coincidences involving physical and astronomical constants have attracted the attention of cosmologists, among others, in recent years and speculation continues regarding what they might mean.

CHAPTER

4

Inverse Probability and the Reverend Thomas Bayes

A form of reasoning that has received a great deal of attention from philosophers, mathematicians, and psychologists is based on theorems proposed by an 18th-century British cleric Thomas Bayes (1763) and the French astronomer and mathematician Pierre Laplace (1774). Today Bayes's name is more strongly associated with this development than is that of Laplace, who is better known for his work on celestial mechanics culminating in a five-volume opus by that title and his treatise on probability theory published in 1812; but Todhunter (1865/2001) credits Laplace with being the first to enunciate distinctly the principle for estimating the probability of causes from the observations of events. Although the popularity of Bayesian statistics waxed and waned—perhaps waned more than waxed—over the years, it has enjoyed something of a revival of interest among researchers during the recent past, as indicated by a near doubling of the annual number of published papers making use of it during the 1990s (Malakoff, 1999).

In classical logic, given the premise "If P, then Q," one cannot argue from the observation Q to the conclusion P; such an argument is known as "affirmation of the consequent" and is generally considered fallacious. Nevertheless both in everyday life and in scientific reasoning, we do take the observation of Q as evidence favoring the likelihood of P: If my hypothesis is correct (P), mixing these two chemicals should produce a small explosion (Q); observing a small explosion upon mixing the chemicals (Q) strengthens my confidence in the correctness of my hypothesis. Most of us would probably agree, perhaps with certain caveats, that that is as it should be.

Although Bayes is remembered for having quantified this idea, the idea itself, or something close to it, was probably around long before Bayes. McLeish (1994) sees the logic that underlies Bayes's rule in the reasoning represented in the Talmudic law, as, for example, when a decision must be made as to whether a child of a recently widowed and remarried woman is more likely that of the deceased husband or the new one. Inasmuch as the baby was born 9 months after the death of the woman's first husband and 6 months after her remarriage, either paternity is possible. The reasoning takes account of the probabilities of full-term and premature births, as well as that of a woman showing signs of pregnancy within 3 months of conception.

THE THEORY

Bayes formalized his notion in what has sometimes been referred to as the "inverse probability theorem," which may be represented as follows:

$$p(H_i|D) = \frac{(D|H_i)p(H_i)}{\sum_{k=1}^{n} p(D|H_k)p(H_k)},$$

where $p(H_i | D)$ is the probability of Hypothesis i given the datum D (the posterior probability of H_i), $p(D | H_i)$ is the probability of the datum D given Hypothesis i (the probability of D conditional on H_i), $p(H_i)$ is the prior probability of Hypothesis i (the probability of H_i before the observed datum D), and n is the total number of hypotheses in the set. Because

$$\sum_{k=1}^{n} p(D|H_k)p(H_k) = p(D),$$

the equation can be simplified as

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)}.$$

Sometimes, to make more explicit the interest in justifying reasoning from effects to causes that motivated the original proposal of the inverse probability theorem, it is expressed as

$$p(C \mid E) = \frac{p(E \mid C)p(C)}{p(E)},$$

where *C* and *E* represent cause and effect, respectively. Conceptualizing the relationship in terms of hypotheses and data gives it a somewhat broader connotation than does conceptualizing it in terms of causes and effects, because the hypotheses can include, but are not necessarily restricted to, causal hypotheses and, similarly, effects are one type of data that can figure in Bayesian reasoning but not the only one.

From Conditional Probabilities to Conditional Probabilities

Bayes's rule is nothing more nor less than a formula for computing a specific probability *given* that specific other probabilities are known. It tells us how to compute the probability of a hypothesis conditional on some data, *if* we already know the probability of that hypothesis and all competing hypotheses in the absence of the data *and,* for each hypothesis, the probability of observing the data in question conditional on that hypothesis. In short, one has to know quite a lot about a situation in order to apply Bayes's theorem, or, if one does not know the essential probabilities, one must make some assumptions or guesses about them. Much of the contention about the application of Bayes's theorem in particular cases has to do with the question of the justifiability of assumptions or guesses that are made.

The rule may be applied iteratively to accommodate a sequence of observations, in which case the value of $p(H_i \mid D)$ that results from any given computation becomes the $p(H_i)$ for the next iteration. Thus the following representation is appropriate for the situation in which Bayes's theorem is to be used when updating a posterior probability estimate with a sequence of observations:

$$p(H_i \mid D_n) = \frac{p(D_n \mid H_i)p(H_i \mid D_{n-1})}{p(D_n)},$$

where $p(H_i \mid D_n)$ represents $p(H_i)$ after the *n*th observation. To calculate $p(H_i \mid D_1)$, the probability of H_i following the first observation, $p(H_i \mid D_{n-1})$ in the equation would be replaced with $p(H_i)$, which represents the a priori, or prior, probability of H_i—the probability of H_i before any observations are made.

Likelihood Ratio

A concept that is often encountered in the Bayesian literature is that of the *likelihood ratio.* A likelihood ratio is the ratio of two conditional probabilities.

$$L_{i,j} = \frac{p(D|H_i)}{p(D|H_j)},$$

for example, is said to be the likelihood ratio of D given H_i to D given H_j. People sometimes find it easier to estimate likelihood ratios than to estimate the underlying conditional probabilities, and likelihood ratio estimates can suffice for the application of Bayes's rule in some instances.

The likelihood ratio tells us nothing about the absolute sizes of the conditional probabilities involved, but only about the size of one relative to the size of the other. It does give a good indication, however, of how discriminating a bit of evidence is with respect to the two hypotheses involved. When the ratio is close to 1, the observation does not help much in choosing between them; however, the more it deviates from 1 in either direction, the greater the credence it gives to one hypothesis or the other. This feature of the ratio is usually said to reflect the *diagnosticity* of the data. The greater the diagnosticity of any datum, the more useful it is in helping us choose among candidate hypotheses.

Odds

Another important ratio in Bayesian reasoning is the ratio of the probabilities of two hypotheses, which is usually referred to as the *odds,* and represented as

$$\Omega_{i,j} = \frac{p(H_i)}{p(H_j)}.$$

By Bayes's rule, the odds following the *n*th observation is given by

$$\Omega_{n;i,j} = \frac{p(H_i|D_n)}{p(H_j|D_n)}$$

$$= \frac{\dfrac{p(D_n|H_i)p(H_i|D_{n-1})}{p(D_n)}}{\dfrac{p(D_n|H_j)p(H_j|D_{n-1})}{p(D_n)}}.$$

Inasmuch as the last expression can be rewritten as

$$\left(\frac{p(D_n|H_i)}{p(D_n|H_j)}\right)\left(\frac{p(H_i|D_{n-1})}{p(H_j|D_{n-1})}\right),$$

we have

$$\Omega_{n;i,j} = L_{i,j}\,\Omega_{n-l;i,j}.$$

If there is no ambiguity about the hypotheses involved, we may drop the subscripts representing them and write simply

$$\Omega_n = L\Omega_{n-l}.$$

Thus, to revise the odds so as to take into account a new observation, one multiplies the existing odds by the likelihood ratio of the observation. In other words, the odds for two hypotheses following the *n*th observation is simply the product of the odds following the $(n-1)$th observation and the likelihood ratio of the *n*th observation. So if one knows, or can estimate, the prior odds and the likelihood ratio for a given observation, one can update the odds in light of the observation without estimating conditional probabilities.

C. R. Peterson and Phillips (1966) suggested modifying the preceding equation by adding an exponent on the likelihood ratio,

$$\Omega_n = L^k\Omega_{n-1},$$

so as to make the equation sufficiently flexible to be descriptive of actual performance. When $k = 1$, the equation represents Bayes's rule; $k < 1$ indicates conservative use of data (giving less weight to data than Bayes's rule deems appropriate) and $k > 1$ indicates overweighting of data. As will be noted in chapter 11, the more usual finding has been that of conservatism.

An essential aspect of many normative models of decision making under uncertainty is the identification of an exhaustive and mutually exclusive set of hypotheses about the possible states of the world. Bayes's rule is a formal means of revising the probabilities associated with such a set in the light of new evidence. It states, in a quantitative way, the effect that an observation should have on the strength of each of a set of hypotheses. Its use presupposes a set of exhaustive and mutually exclusive hypotheses, H_i, regarding the state of the world. Exhaustive and mutually exclusive means that the set includes the true hypothesis and that only one of the hypotheses is true. With each hypothesis there is associated a nonzero probability, and these probabilities must sum to one.

Assuming that one has identified an appropriate hypothesis set and has assigned to each hypothesis a prior probability of its being correct, what one then must do in order to use the rule to extract information from an observation is estimate for each of the hypotheses the probability of the observation conditional on that hypothesis, $p(D \mid H_i)$. That is to say, one must estimate the probability that that particular observation would be made, assuming that particular hypothesis is true. With these estimated conditional probabilities in hand, application of the rule is straightforward. The result is a set of posterior probabilities, which may be thought of as the prior probabilities updated to take into account the information extracted from the observation.

AN APPLICATION OF BAYES'S RULE

The following simple thought experiment illustrates how the process works. Imagine three urns, each filled with black and white balls but in different proportions. Specifically, suppose that the balls are 10%, 50%, and 90% black in urns 1, 2, and 3, respectively. Now suppose that you are told that one of the three urns is to be selected at random, and balls will be drawn from it, one at a time, and replaced after each drawing. You will be told the color of the ball that was drawn on each occasion, and your task is to determine from which urn the balls are being drawn.

This problem is ideally suited to a Bayesian analysis. There are obviously three a priori hypotheses, and they constitute an exhaustive and mutually exclusive set. Inasmuch as we have no reason to consider the selection of any one of the urns to be more probable than that of any of the others, it makes sense to assign .333 as the prior probability of each hypothesis. The conditional probabilities are also straightforward. The probabilities of observing a black ball are .1, .5, and .9 for hypotheses 1, 2, and 3, respectively, and of course the probabilities of observing a white ball are the complementary set, .9, .5, and .1. So, starting with the belief that each hypothesis is equally probable, suppose the first draw produces a black ball. What effect does this observation have on the strength of our various hypotheses?

Setting $p(H) = .333$ for all three hypotheses, $p(D \mid H_i) = .1, .5$, and $.9$ for hypotheses 1 through 3, respectively, we apply the formula,

$$p(H_i \mid D) = \frac{p(D \mid H_i)p(H_i)}{\sum_{j=1}^{n} p(D \mid H_j)p(H_j)},$$

which, since the datum in hand is a black ball and there are three hypotheses to consider, we will rewrite as

$$p(H_i \mid B) = \frac{p(B \mid H_i)p(H_i)}{\sum_{j=1}^{3} p(B \mid H_j)p(H_j)}.$$

We get, for the denominator,

$$\sum_{j=1}^{3} p(B \mid H_j)p(H_j) = (.1)(.333) + (.5)(.333) + (.9)(.333) = .5,$$

so

$$p\ (H_1 \mid B) = (.1)(.333)/.5 = .067,$$

$$p\ (H_2 \mid B) = (.5)(.333)/.5 = .333,$$

and

$$p\ (H_3 \mid B) = (.9)(.333)/.5 = .600.$$

The reader who is surprised to see that the posterior probability of the third hypothesis is almost 10 times as large as the posterior probability of the first hypothesis after the drawing of only a single ball is in good company. Most people do not change their estimates as much as Bayes's rule indicates they should as a consequence of observations; this is the well-known finding of conservatism in evidence evaluation that many investigators have described, about which more later. Suppose that we draw a second ball and it too is black. The denominator of our equation is now

$$\sum_{j=1}^{3} p(D_2 \mid H_j)p(H_j \mid B) = (.1)(.067) + (.5)(.333) + (.9)(.600) = .713,$$

so the posterior probabilities after the second observation then would be as follows:

$$p\ (H_1 \mid B,B) = (.1)(.067)/(.713) = .009,$$

$$p\ (H_2 \mid B,B) = (.5)(.333)/(.713) = .234,$$

$$p\ (H_3 \mid B,B) = (.9)(.600)/(.713) = .757.$$

After the drawing of two black balls in a row the probability of the third hypothesis is 84 times as great as the probability of the first. It should be clear that if the first two balls to be drawn were white, the posterior probability of the Hypothesis 1 would be .757 after the second draw and that of Hypothesis 3 would be .009. Hypothesis 2 would be .234 in both cases. Of course, if instead of using urns with proportions of balls of one color as widely disparate as .1, .5, and .9, we used proportions like .48, .50., and .52, say, the effect of a small sample of draws would be much less dramatic.

Now imagine that we draw first a black ball and then a white one. After the first draw—of a black ball—the distribution of posterior probabilities will be as in the first example, .067, .333, and .600 for Hypotheses 1, 2, and 3, respectively. After the second draw, this time of a white ball, they will be as follows:

$$p(H_1 \mid B, W) = (.9)(.067)/(.287) = .210,$$

$$p(H_2 \mid B, W) = (.5)(.333)/(.287) = .580,$$

$$p(H_3 \mid B, W) = (.1)(.600)/(.287) = .210.$$

In keeping with our intuitions, drawing a ball of each color on two successive draws increases the posterior probability that the urn from which the balls are being drawn is the one containing black and white balls in equal numbers, and decreases the posterior probabilities of the competing hypotheses equally, but not by enormous amounts. The reader will easily verify that the distribution of probabilities following the drawing of a ball of each color would be the same independently of whether the black or the white ball was drawn first. If this were not the case, the outcome would so strongly violate our intuitions that we would find it hard to take the rule seriously.

This example is very simple, in that there is no difficulty in determining what the set of hypotheses should be and there is no problem in assigning values for either prior or conditional probabilities. Most real-life problems are not this straightforward. Also there is really no need for human judgment in this example. However, it is interesting to consider how closely the posterior probabilities that people assign on an intuitive basis match those produced by application of Bayes's rule even in comparably simple cases. Because of experimental results showing people to be too conservative in their estimates of posterior probabilities (see later discussion) some investigators have proposed that people should be relied upon to make conditional probability judgments and that posterior probabilities should be calculated—from these estimated conditionals—automatically.

A difficulty with this view is that when the conditional probabilities are known, as in the previous example, there is no need for anyone to estimate them, and when they are unknown it is not clear how accurate people's estimates of them are. In many instances, there is no alternative to estimates produced by people. In such cases, people could be asked to estimate either conditional or posterior probabilities. The opinion that people should be asked to estimate conditionals from which posteriors can be calculated rests on the assumption that the greater ability of people to estimate conditionals in situations in which they can be determined objectively carries over to situations in which they cannot be so determined.

A tacit assumption throughout the foregoing discussion has been that the observations—the data—used to update posterior probabilities are unambiguous. The ball that is drawn from the urn is either black or white, and upon drawing and observing it there is no question as to which is the case. Often in real life the evidence to which one has access is not so assuredly reliable. The situation is more like that in which one gets information through a noisy channel; one does not observe the ball directly, for example, but gets a report from an observer who sometimes lies or makes a mistake.

Techniques have been proposed for taking account of the unreliability of data in applying Bayes's rule (Gettys & Wilke, 1969; Schum & DuCharme, 1971; Schum, DuCharme, & DePitts, 1973; Snapper & Fryback, 1971; Steiger & Gettys, 1972). They all involve introducing one or more terms in the computation to adjust the magnitude of the influence that a reported observation can have in updating the probabilities of alternative hypotheses and they have the effect of decreasing the diagnosticity of reported data. We will not consider these techniques here, but it is important to understand that the need for them arises because of the fact that Bayes's rule assumes reliable data and that assumption often is not valid in real-life situations.

BAYESIAN DECISION MAKING

Structured approaches to decision making often require identification of the possible states of the world, estimation of the probability of each possibility, identification of the set of actions open to the decision maker, specification of the outcome expected for each possible state–action pairing, the value (worth, utility) of that outcome, and the application of a rule that selects the course of action in accordance with a specific goal, such as maximization of expected utility. The only part of this process that is distinctively Bayesian, in the sense that it makes use of Bayes's rule, is the estimation of the probabilities associated with the possible states of the world, and even here it involves only the revision of these probabilities in the light of newly acquired data; although it requires that there be an exhaustive and mutually exclusive set of such possibilities, it has nothing to contribute to the identification of the possibilities or to the assigning of a priori probabilities to them.

In short, Bayes's rule is intended to help one determine, in the light of data, how much one should revise one's beliefs about the hypotheses being entertained. It does not tell one what those hypotheses should be in the first place or what one should do, given any particular belief set. One may decide to act as though a particular hypothesis were true if an analysis yields a posterior probability or likelihood ratio that exceeds some criterion value, but that decision strategy is not dictated by Bayes's rule. And the criterion value should be determined by taking into account not only the benefit of acting as though the hypothesis were true if it really is true, but also the consequences of other possible outcomes, such as the cost of acting as though the hypothesis were true when it really is not and the cost of acting as though it were false when it really is true.

Although Bayes's rule has the very limited role just described in the context of decision making, the term "Bayesian decision making" has come to have a considerably broader connotation, and sometimes is used to refer to the entire decision process of which the updating of hypothesis probabilities is only a part. Unfortunately, the term has lost a lot of precision as a consequence of this broadening. Good (1983a, p. 20) once pointed out that the varieties of philosophies that have been called Bayesian exceeds the number of professional statisticians. He calculated the number of varieties (46,656) as all possible combinations of 11 issues with respect to which Bayesian statisticians differ. One conclusion that he drew from this observation is that some of the arguments that are advanced against *the* Bayesian position are valid only against *some* Bayesian position(s). Another implication of the observation is that to say that one is a Bayesian is to reveal very little regarding what one believes about reasoning under uncertainty.

Expository treatments of Bayesian decision making have been given by Edwards, Lindman, and L. J. Savage (1963), Lindley (1965, 1971), Slovic and Lichtenstein (1971), Phillips (1973), Novick and Jackson (1974), Fischhoff and Beyth-Marom (1983), R. Jeffrey (1983), Howson and Urbach (1993), and Hacking (2001), among many others.

THE JUSTIFIABILITY OF BAYES'S RULE

> The deduction of the probability of causes from the probability of their consequences is a game whose rules are such that no one can take part in it without cheating. (Eddington, 1935, p. 126)

The question of whether Bayes's rule is descriptive of how people actually apply new data to the modification of their beliefs has received a great deal of attention from researchers, and we will consider some of the experimental results presently. Whether or not people behave like Bayesians, Bayes's rule has been promoted widely as a normative, or prescriptive, theory of how beliefs *should* be changed as a consequence of the acquisition of new data that relate to them. Not everyone who has thought about it has readily accepted this view, however, and debates about what inverse probability, which is what the rule is said to compute, really means go back to the time of Bayes.

Where Does One Find Prior Probabilities?

A problem perceived by Bayes (who never published his rule) was the need to estimate prior probabilities. In order to apply the theorem, quantitative values for prior probabilities must be assigned to the competing hypotheses under consideration. In many practical situations, there is no objective way to determine prior probabilities, so if the rule is to be applied in these cases, subjective probabilities must be used. In the absence of information that would justify some other assumption, there seems to be no alternative to assuming that all hypotheses are equally likely a priori. Today this might be justified on the grounds that it is the assumption that maximizes a priori uncertainty, in the information-theoretic sense, and this seems appropriate for the situation in which one has no basis for considering any hypothesis to be more or less likely than any other. Bayes had trouble with the "uniform prior probability distribution" assumption, whereas Laplace did not.

One of the implications of the use of subjective estimates of prior probability distributions, when agreed-upon objective distributions are not known, is that different individuals applying Bayes's rule to the same observations will produce different posterior probabilities if their subjective estimates of the prior probabilities differ. To people who consider probability to be a subjective entity in any case, this poses no problem, because on the view that a probability reflects nothing more than a state of mind—a degree of knowledge or ignorance of the situation of interest—there is little reason to assume equality among all observers in that regard. To one who wishes to give probability a more objective meaning, this feature of Bayes's rule is bothersome, and the general applicability of the rule has been challenged by holders of a frequentistic view of probability (Venn, 1888; von Mises, 1928/1957). Eddington (1935) expressed a negative view of the use of subjective probabilities in scientific calculations this way: "the most elementary use of the word probability refers to strength of expectation or belief which is not associated with any numerical measure. There can be no exact science of these non-numerical probabilities which reflect personal judgment" (p. 113). Bunge (1996), criticizes Bayesians for using probability—"or rather their own version of it"—when confronted with ignorance or uncertainty: "This allows them to assign prior probabilities to facts and propositions in an arbitrary manner—which is a way of passing off mere intuition, hunch, guess for scientific hypothesis" (p. 103).

Sensitivity to Prior Probabilities

In view of the controversial status of the need, at least sometimes, to assign subjective estimates to initial prior probabilities, we would like to know how sensitive the Bayesian approach is to the values that the prior probabilities are given. If the effect of the distribution of prior probabilities is invariably very small compared to the effect of the application of the rule to new data, then one would not have to worry much about the accuracy of the prior probability estimates, at least for situations in which many observations, and updates to the posterior distribution, are to be made.

We can get some feel for this issue by considering again the problem of deciding from which of three urns balls are being drawn. Suppose that, for whatever reason, we started out believing that one of the urns, say the one with black and white balls in the ratio 1 to 9, was very likely to be the one from which the sample was to be drawn, so we assigned a prior probability of .8 to this urn and .1 to each of the others. What would the posterior probabilities be after the drawing of a single black ball?

Setting $p(H_1)$ at .8 and $p(H_2)$ and $p(H_3)$ both at .1, and again letting $p(D \mid H_1)$ = .1, .5, and .9 for Hypotheses 1 through 3 respectively, we get, for the denominator of our Bayesian formula

$$\sum_{j=1}^{3} p\left(D \middle| H_j\right) p\left(H_j\right) = (.1)(.8) + (.5)(.1) + (.9)(.1) = .22,$$

so

$$p(H_1 \mid B) = (.1)(.8)/.22 = .364,$$

$$p(H_2 \mid B) = (.5)(.1)/.22 = .227,$$

and

$$p(H_3 \mid B) = (.9)(.1)/.22 = .409.$$

Now suppose we draw a second black ball

$$\sum_{j=1}^{3} p\left(D_2 \middle| H_j\right) p\left(H_j \middle| D_1\right) = (.1)(.364) + (.5)(.227) + (.9)(.409) = .518,$$

so the posterior probabilities after the second observation would be:

$$p(H_1 \mid B,B) = (.1)(.364)/(.518) = .070,$$

$$p(H_2 \mid B,B) = (.5)(.227)/(.518) = .219,$$

$$p(H_3 \mid B,B) = (.9)(.409)/(.518) = .711.$$

Thus the distribution of probabilities after the drawing of two successive black balls is almost the same when we start with the a priori distribution of .8, .1, .1 as when we start with the distribution .333, .333, .333. The full situation is laid out in Table 4.1.

This illustration does not prove the point, of course, but it may make plausible the idea that given only a few observations, Bayes's rule is relatively insensitive to prior probabilities, provided they are not very close to 0 or 1. The reader may wish to experiment with various combinations of prior probability distributions and observation sequences to get a better feel for how probability distributions change in response to the acquisition of data.

Let us suppose, for the sake of discussion, that the problem of having to assign initial prior probabilities subjectively is not a serious one when the situation is such that the initial estimates are to be updated many times as the consequence of a series of independent observations, as in the example with the drawing of balls from an urn, because a few observations will wash out the effects of all but very extreme initial distributions in any case. Unfortunately,

TABLE 4.1

Posterior Probabilities of Each of Three Hypotheses About the Proportion of Black Balls in an Urn After *n* Draws, Each of Which Produces a Black Ball

	Case 1			*Case 2*		
n	*0*	*1*	*2*	*0*	*1*	*2*
$P(H_1 \mid D_n)$	*.333*	*.067*	*.009*	*.8*	*.364*	*.070*
$P(H_2 \mid D_n)$	*.333*	*.333*	*.234*	*.1*	*.227*	*.219*
$P(H_3 \mid D_n)$	*.333*	*.600*	*.727*	*.1*	*.409*	*.711*

Note. H_1: proportion of black balls = .1; H_2: proportion of black balls = .5; H_3: proportion of black balls = .9. In Case 1, the prior probabilities for the three hypotheses are equal at .333; in Case 2, they are .8, .1, and .1 for Hypotheses 1, 2, and 3, respectively.

any comfort we derive from this supposition does not apply to situations in which only a single observation is to be made. There are many real-life, or real-life-like, situations in which one has the results of a single observation and wants to know what to make of them. A relatively reliable, but not infallible, medical test has proved positive; what do I conclude from that regarding the probability that I have the rare disease for which the test is diagnostic? An acquaintance in whose judgment I have considerable, but not total, confidence tells me that ... ; how seriously should I take this claim? Answering these questions within a Bayesian framework means revising prior probabilities in the light of the observations, and inasmuch as there is only one observation in each case, how the prior probabilities are estimated and what the estimates are are of some importance.

BASE RATES AND PRIOR PROBABILITIES

Sometimes, as in the case of the ball-sampling problem described previously, the appropriate prior probabilities are clear; often, however, they are not. A conventional basis for estimating these probabilities, when the information is available, is some measure of *base rate,* for example, the incidence of a disease of interest in the population. The appropriateness of the use of base rates for this purpose has been the subject of heated debate, although perhaps it is more accurate to say that the argument has had less to do with whether base rates should be used in this way than with the question of *what* base rates are appropriate in specific cases. The point is illustrated by reference to an article by L. J. Cohen (1981), on the question of whether human irrationality can be experimentally demonstrated, that was published along with commentary by several other experts on the subject. The article and commentary are worth considering in some detail, because they illustrate the considerable differences that can exist among the opinions of highly knowledgeable people regarding what constitutes rational behavior in a given situation.

The Cab-Color Problem

L. J. Cohen (1981) focuses on two situations that are typical of those to which Bayesian analysis is usually considered appropriate. The first is a problem attributed to Kahneman and Tversky (1972a) that goes as follows. Eighty-five percent of the taxi cabs in a given city are blue, and the remaining 15% are green. A cab that was involved in a hit-and-run accident at night was identified by a witness as green. Tests showed that the witness was able to distinguish a blue cab from a green cab under night lighting conditions four out of five times.

What is the probability that the cab involved in the accident was blue? The median probability estimate produced by the participants in the cited experiment was .20. This result suggested to the investigators that the participants ignored prior probabilities and based their estimates entirely on what they were told regarding the accuracy of the witness in identifying cab color.

According to one view, the correct answer is given by the formula

$$p(B\mid g)=\frac{p(g\mid B)p(B)}{p(g\mid B)p(B)+p(g\mid G)p(G)},$$

where $p(B \mid g)$ is the posterior probability that the cab is Blue given that the witness said "green," $p(g \mid B)$ is the probability that the witness would say "green" given that the cab was Blue, and $p(B)$ is the prior probability that the cab was Blue. The prior probability distribution represents what one should consider the probabilities to be if one did not have the testimony of a witness.

In fact, given the aforementioned statement of the problem, we cannot solve this equation exactly because we do not know the value of $p(g \mid B)$ or $p(g \mid G)$. The problem statement tells us only that the witness could identify the correct cab color four times out of five, but it does not tell us the conditional probabilities of the two types of errors: saying "blue" when the cab was Green and saying "green" when it was Blue. We will return to this point later. For now, if we make the assumption that these conditional probabilities were equally likely, and we use the 85-to-15 ratio of Blue-to-Green cabs as the basis for assigning prior probabilities, $p(B) = .85$ and $p(G) = .15$, then the computation is:

$$p(B \mid g) = (.2)(.85)/[(.2)(.85)+(.8)(.15)] = .17/.29 = .59.$$

According to this computation, the posterior probabilities estimated by participants in the cited study were low by a factor of about three.

In some experiments in which the same problem has been used, the experimenters have specified in the problem description that the conditional probabilities, $p(g \mid B)$ and $p(b \mid G)$, were equally likely. Lyon and Slovic (1976), for example, gave their participants the information about the witness's accuracy this way: "The court tested the witness's ability to distinguish a blue cab from a green cab at night by presenting to him film sequences, half of which depicted blue cabs, and half depicting green cabs. He was able to make correct identification in eight out of ten tries. He made one error on each color of cab" (p. 291). Lyon and Slovic's results were very similar to those of Kahneman and Tversky (1972) in showing people's estimates to be quite insensitive to prior probabilities, as represented by the relative number of cabs of each color in operation.

The situation may be represented in abstract form as in Table 4.2, the cells of which are probabilities that sum to 1.0.

There are several probabilities represented by this matrix, and it is important to keep the distinctions among them in mind. A, B, C and D are joint probabilities:

A: $p(B\&b)$, the probability the cab is Blue *and* the witness says "blue."

B: $p(G\&b)$, the probability the cab is Green *and* the witness says "blue."

C: $p(B\&g)$, the probability the cab is Blue *and* the witness says "green."

D: $p(G\&g)$, the probability the cab is Green *and* the witness says "green."

E, the sum of $p(B\&b)$ and $p(B\&g)$, is the probability that a randomly selected cab is Blue and F, the sum of $p(G\&b)$ and $p(G\&g)$, is the probability that it is Green. G is the probability that the witness says "blue" and H the probability that he says "green." Of greatest interest for present purposes are the following conditional probabilities:

A/E: $p(b \mid B)$, the probability the witness is correct *given* the cab is Blue.

D/F: $p(g \mid G)$, the probability the witness is correct *given* the cab is Green.

A/G: $p(B \mid b)$, the probability the witness is correct *given* he says "blue."

D/H: $p(G \mid g)$, the probability the witness is correct *given* he says "green."

TABLE 4.2

Representation of the Cab Colors—Witness Report Problem

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	A	B	G
	"green"	C	D	H
	Column Σ	E	F	1.00

Note. Each cell entry, A, B, C and D, represents the joint probability of the cab being the color indicated by the associated column and the witness reporting the color indicated by the associated row. Marginal entries, E, F, G, and H, are row and column sums. (Note that B and G, as markers in the table, are not to be confused with *B* and *G* standing for Blue and Green cabs in probability notation in the text.)

Note that the unconditional probability that the witness makes a correct identification is A + D, the probability that the cab is Blue *and* the witness says "blue" plus the probability that the cab is Green *and* the witness says "green."

Table 4.3 shows the situation defined by the conditions of Lyon and Slovic's (1976) study, in which the prior probabilities of the cab being Blue and Green were .85 and .15, respectively, and the probability that the witness's report was correct was .8 irrespective of the cab's actual color. The conditional probabilities of interest are represented here only implicitly; the probability that the witness says "green" given that the car is actually Blue, $p(g \mid B)$, for example, is the ratio of the joint probability of the car being Blue and the witness saying "green," $p(B\&g)$, to the probability of the car being Blue, $p(B\&g) + p(B\&b)$, which is .17/.85, or .2.

According to the conventional Bayesian approach to this problem, the probability that the cab was actually Blue, given that the witness said it was "green," is computed from the application of Bayes's rule, using the base rates .85 and .15 as the prior probabilities of the cab being blue and green, respectively, and applying the conditional probabilities of the witness to derive the posterior probabilities, which represent the probability of the cab being a specified color that takes both the prior probabilities and the observational data into account. The computation, as was noted previously, gives .17/.29, or .59, as the posterior probability that the cab was blue, given that the witness reported it to be green. The situation is as shown in Table 4.3.

L. J. Cohen (1981) argues that .59 is not the correct answer to the question. What this ratio represents, in his words, is as follows:

> [It represents] the value of the conditional probability that a cab-colour identification by the witness is incorrect, on the condition that it is an identification as green. Jurors, however, or people thinking of themselves as jurors, ought not to

TABLE 4.3

The Analysis Shown in Table 4.2 With the Values Used in the Taxicab Problem Discussed Earlier

		Cab Color		
		Blue	*Green*	
Witness Says	"blue"	*.68*	*.03*	*.71*
	"green"	*.17*	*.12*	*.29*
		.85	*.15*	*1.00*

> rely on that probability if they can avoid doing so, since reliance on it assumes the issue before the court to concern a long run of cab-colour identification problems—whereas in fact it concerns just one problem of this type. Jurors here are occupied, strictly speaking, just with the probability that the cab actually involved in the accident was blue, on the condition that the witness said it was green. And the latter probability is equivalent in the circumstances to the probability that a statement to the effect that the cab actually involved in the accident was green, is false, on the condition that the statement is made by the witness. If the jurors know that only 20% of the witness's statements about cab colours are false, they rightly estimate the probability at issue as 1/5 without any transgression of Bayes's law. (p. 328)

Before turning to the second situation that L. J. Cohen (1981) considered, I want to expand on the point that was made earlier: that the first problem, as stated, does not have an unequivocal answer, because of the ambiguity of the claim that the witness can distinguish blue cabs from green ones in 80% of cases. What might this claim mean? Presumably it means that of all the responses the witness makes when asked to identify the color of a cab that is either blue or green, 80% are correct. What this does not tell us is whether, when the witness makes a mistake, he is as likely to misidentify a blue cab as a green one. And this is critical. Consider the three possibilities outlined in Tables 4.4, 4.5, and 4.6.

In all three cases, we interpret the claim that the witness can distinguish blue cabs from green in 80% of instances to mean that, when presented with cabs at random in such a way that the probability of a cab being blue is .5 on each trial, he identifies the colors correctly on 80% of the trials. In Case 1, we assume further that the witness is as likely to make the one kind of mistake as the other. In Case 2, we assume that he frequently misidentifies a Green cab as "blue," but never misidentifies a Blue one as "green." In Case 3, we make the opposite assumption. In all three cases the probability of a correct identification, all trials

TABLE 4.4
Case 1

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	*.40*	*.10*	*.50*
	"green"	*.10*	*.40*	*.50*
	Column Σ	*.50*	*.50*	*1.00*

TABLE 4.5
Case 2

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	.50	.20	.70
	"green"	.00	.30	.30
	Column Σ	.50	.50	*1.00*

TABLE 4.6
Case 3

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	.30	.00	.30
	"green"	.20	.50	.70
	Column Σ	.50	.50	1.00

combined, is .8, but only in Case 1 is it true that the probability that the witness is correct is .8 independently of whether he says "green" or "blue." The probability that he is correct, given that he says "green," is 1.0 and .71 in Cases 2 and 3, respectively; the probability that he is correct, given that he says "blue," is .71 and 1.0 in the same two cases.

If we plug the appropriate conditional probabilities from these matrices into the Bayesian calculation and use .85 as the prior probability that the cab is blue, we get as the posterior probability, in the three cases:

Case 1: $p(B \mid g) = (.2)(.85)/[(.2)(.85) + (.8)(.15)] = .17/.29 = .59.$

Case 2: $p(B \mid g) = (0)(.85)/[(0)(.85) + (.6)(.15)] = 0.0.$

Case 3: $p(B \mid g) = (.4)(.85)/[(.4)(.85) + (1.0)(.15)] = .34/.49 = .69.$

Alternatively, if we accept Cohen's argument that "the fact that cab colours actually vary according to an 85/15 ratio is strictly irrelevant" to the estimate of the probability that the cab is Blue, given that the witness says it is "green," and use .5 as the prior probability, we get for the three cases, again using Bayes's rule:

Case 1: $p(B \mid g) = (.2)(.5)/[(.2)(.5) + (.8)(.5)] = .1/.5 = .2$.

Case 2: $p(B \mid g) = (0)(.5)/[(0)(.5) + (.6)(.5)] = 0.0$.

Case 3: $p(B \mid g) = (.4)(.5)/[(.4)(.5) + (1.0)(.5)] = .2/.7 = .29$.

L. J. Cohen (1981) argues that .2 is the correct answer, not because it comes from the application of Bayes's rule, as in Case 1 as just listed, but on the basis of its being the conditional probability of the witness saying "green" when the cab is really Blue. As we have noted, this conditional probability is an assumption that goes beyond what is given in the statement of the problem. Cohen undoubtedly realized the need for this assumption, because he mentioned that Lyon and Slovic (1976) had made it explicit in an article that postdated Kahneman and Tversky's original one.

Another conceivable, though perhaps less likely, interpretation of the claim that the witness distinguishes blue cabs from green in 80% of cases is that he does so when he encounters the two colors with the relative frequency with which cabs of these colors appear around town, which is to say when 85% of the instances he has to judge are blue and 15% are green. With this interpretation, we again can imagine three cases analogous to those described previously (see Tables 4.7, 4.8, and 4.9).

(Note: In Cases 5 and 6, I used the largest probabilities in the Blue *and* "blue" and Green *and* "green" cells, respectively, that were possible with the constraint that the overall probability of a correct identification stay constant at .8. In Case 6, it was possible to have the probability of the Green *and* "green" cell match the probability of the occurrence of a Green cab, thus making $p(\text{correct} \mid \text{Green}) = 1.0$. In Case 5, it was not possible to have the probability of Blue *and* "blue" equal the probability of the occurrence of a Blue cab, because any-

TABLE 4.7
Case 4

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	.68	.03	.71
	"green"	.17	.12	.29
	Column Σ	.85	.15	1.00

TABLE 4.8

Case 5

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	.80	.15	.95
	"green"	.05	.00	.05
	Column Σ	.85	.15	1.00

TABLE 4.9

Case 6

		Cab Color		
		Blue	*Green*	*Row Σ*
Witness Says	"blue"	.65	.00	.65
	"green"	.20	.15	.35
	Column Σ	.85	.15	1.00

thing greater than .8 in this cell would have made the overall probability of a correct identification be greater than .8.)

Again, in all three cases the probability of a correct identification, all trials combined, is .8, but the probability that the identification is correct conditional on a particular identification response varies with the specifics of the cases. The conditional probabilities of interest are as shown in Table 4.10.

Let us again calculate with Bayes's rule the probability that the cab was Blue, given that the witness reported it to be "green," using the values assumed in these three cases:

Case 4: $p(B \mid g) = (.2)(.85)/[(.2)(.85) + (.8)(.15)] = .17/.29 = .59.$

Case 5: $p(B \mid g) = (.06)(.85)/[(.06)(.85) + (0)(.15)] = .05/.05 = 1.00.$

Case 6: $p(B \mid g) = (.24)(.85)/[(.24)(.85) + (1.0)(.15)] = .2/.35 = .57.$

It comes as no surprise, of course, that these probabilities are the complements of those of the last row in Table 4.10, inasmuch as the probability that the

TABLE 4.10
The Probability of Being Correct, Conditional on Either of the Two Car Colors or Either of the Two Reports, for Cases 4, 5, and 6 Discussed Earlier

	Case		
	4	*5*	*6*
P(correct I Blue)	.80	.94	.76
P(correct I Green)	.80	.00	1.00
P(correct I "blue")	.96	.84	1.00
P(correct I "green")	.41	.00	.43

cab is Blue, given that the witness says "green," is the complement of the probability that the witness is correct, given that he says "green." The correspondence of the numbers simply shows that one gets the same result if one uses Bayes's rule as one does when one obtains the probabilities directly by considering the appropriate outcome ratios.

The cases just considered illustrate that the claim that the witness can distinguish blue cabs from green ones in 80% of cases does not provide the information needed to support an unqualified answer to the question of what is the probability that the cab in the accident was Blue, given that the witness has identified it as "green." One might acknowledge the legitimacy of these various interpretations of the claim that the witness can distinguish blue cabs from green ones in 80% of cases, but argue that the most natural interpretation is one that assumes that the witness is equally as likely to identify a blue cab as green as to identify a green one as blue, and that in the absence of evidence to the contrary, we should assume this to be the one that people naturally make. It seems to me a reasonable working assumption, but it is important to recognize that it is an assumption.

L. J. Cohen (1981) argues that "[t]he fact that cab colours actually vary according to an 85/15 ratio is strictly irrelevant to this estimate [the estimate of the probability that the cab was blue, given that the witness has said it was green], because it neither raises nor lowers the probability of a specific cab-colour identification being correct on the condition that it is an identification by the witness" (p. 328), and he justifies this view on the grounds that "[a] probability that holds uniformly for each of a class of events because it is based on causal properties, such as the physiology of vision, cannot be altered by facts, such as chance distributions, that have no causal efficacy in the individual events" (p. 329).

A conventional counter to this argument is that, given the assumption that the probability that the witness will misidentify the color on a single trial is .2, independently of whether the cab is actually blue or green, then a long run of trials *on 85% of which the cab is blue* will produce the outcome represented by Case 4. If we think of the experiment as a random one from that long run of imaginary trials, and take the theoretical relative frequencies as indicative of the momentary probabilities, then the numbers in the matrix associated with Case 4 represent also the probabilities associated with that trial, and the correct answer to the question of the probability that the cab was blue, given that the witness said it was green is .59. Cohen argues that a relative-frequency conception of probability is not appropriate here, that probability in this case is better viewed as a "causal propensity," the judgment of which does not rely on relative frequencies, and that the sought for probability is .2, the probability that the witness makes an error.

I confess to not fully understanding Cohen's argument. It seems to me he dismisses the relevance of relative frequencies, but uses them all the same. He denies the relevance of the relative frequencies of blue and green cabs in operation, but he accepts the relative frequency with which the witness makes errors of identification (Margalit & Bar-Hillel, 1981). And it is not clear to me how an error of identification qualifies as a causal propensity any more than does the relative numbers of cabs of different colors in operation. One can think of causes of errors of identification (poor vision, poor lighting), but in what sense is the causal connection between such factors and the problem greater than is the relative availability of cabs of specified colors for involvement in an accident?

In defending Cohen's analysis of the taxicab problem, Levi (1981) argues that the percentage of the total population of cabs that are blue (or green) is not relevant to the jurors' task; what they need to be told, he claims, is "the percentage of blue (green) cabs in the city involved in accidents" (p. 343). (See also Niiniluoto [1981].) A good Bayesian, he argues, "should neglect the base rate given in the example because it is useless for the purpose of determining subjective probabilities" (p. 343). He notes too that when people have been told that 85% of the taxicabs involved in accidents have been blue, they have not neglected base rates.

I would agree that knowledge of the percentage of cabs of a specified color that have been involved in accidents should count for more than knowledge of the percentage of cabs of a specified color in the town's fleet. If told, for example, that 75% of the cabs in town are green, but that 75% of the cabs that have been involved in accidents have been blue, I would think the latter percentage a more appropriate basis than the former for estimating the prior probability that a random accident will involve a blue cab. But the question is, if one does not have the accident statistics and has only the population statistics, should one use the latter or should one ignore them. It seems to me one should use them, but that one should have less confidence in one's answer than would be justified if it were based on more situation-specific base rate data.

A Problem of Disease Diagnosis

Let us turn to the second situation considered by L. J. Cohen (1981) in his discussion of the possibility of demonstrating human irrationality. A patient suffers from one of two diseases, A or B. Imagine yourself as that patient. Cohen posits the following:

> For a variety of demographic reasons disease A happens to be nineteen times as common as B. The two diseases are equally fatal if untreated, but it is dangerous to combine the respectively appropriate treatments. Your physician orders a certain test which, through the operation of a fairly well understood causal process, always gives a unique diagnosis in such cases, and this diagnosis has been tried out on equal numbers of A- and B-patients and is known to be correct on 80% of those occasions. The tests report that you are suffering from disease B. Should you nevertheless opt for the treatment appropriate to A, on the supposition (reached by calculating as the experimenters did) that the probability of your suffering from A is 19/23? Or should you opt for the treatment appropriate to B, on the supposition (reached by calculating as the subjects did) that the probability of your suffering from B is 4/5? (p. 329)

Cohen leaves no doubt of his answer to the question: "It is the former option that would be the irrational one for you, qua patient, not the latter" (p. 329).

Note first that, as was true of the taxicab problem as presented, the description of this problem is, strictly speaking, inadequate to permit unequivocal computation of posterior probabilities, whatever one's attitude toward base rates. The problem is with the claim that the diagnosis "has been tried out on equal numbers of A- and B-patients and is known to be correct on 80% of those occasions" (p. 329). The statement is sufficiently vague to permit the interpretation that 80% of the total number of diagnoses were correct without necessarily assuming that 80% of the A-patients were diagnosed correctly and that 80% of the B-patients were diagnosed correctly. The statement does not rule out the possibility, for example, of an overall rate of 80% correct diagnoses, with 90% of one type of patient being diagnosed correctly and 70% of the other being so. As in the case of the taxicab scenario, one might be willing to assume that most readers would interpret the claim as that of equal accuracy of diagnosis for both types of patient, but it is important to recognize that that is an assumption and it is essential to the further analysis of the problem. I believe that a nontrivial amount of the difficulty people have in dealing with questions of probability stem from the imprecise use of language, and it therefore is better to err in the direction of more-than-adequate precision than in the opposite one. In any case, for the sake of further discussion, let us take the intent of the claim to be that the probability of a misdiagnosis given the presence of disease A was the same as that of a misdiagnosis given the presence of disease B and that it was .2 in both cases. Given this interpretation, we may represent the situation described by Cohen as in Table 4.11.

The conventional Bayesian answer to the question of the probability that you are suffering from Disease A, given that the tests indicate that you are suffering from Disease B, which I will represent as $p(A \mid b)$, is

$$p(A \mid b) = \frac{p(b \mid A)p(A)}{p(b \mid A)p(A) + p(b \mid B)p(B)},$$

$$p(A \mid b) = (.2)(.95) / [(.2)(.95) + (.8)(.05)] = .19/.23 = .83.$$

As noted earlier, L. J. Cohen (1981) rejects this answer, arguing that the standard statistical method of taking the prior frequency into account would be correct "if what one wanted was a probability for any patient considered not as a concrete particular person, not even as a randomly selected particular person, but simply as an instance of a long run of patients" (p. 329). But the individual patient, concerned only about her own situation, Cohen argues, "needs to evaluate a propensity-type probability, not a frequency-type one, and the standard stastical method would then be inappropriate" (p. 329). Cohen contends that this view is not a repudiation of Bayesian analysis, but only of the assumption

TABLE 4.11

Joint Probability of the Disease State Indicated by the Associated Column and the Diagnosis Indicated by the Associated Row

		Disease State		
		A	*B*	*Row Σ*
Diagnosis	"a"	.76	.01	.77
	"b"	.19	.04	.23
	Column Σ	.95	.05	1.00

that the prior probabilities are appropriately taken to be equivalent, in this case, to the relative frequency of occurrence of the two types of disease in the population. Cohen's analyses of the taxicab and medical-test problems evoked much commentary, positive and negative.

The Issue of Relevance

L. J. Cohen's (1981) argument is that the appropriate application of Bayes' rule requires that the prior probabilities be based on knowledge that is *relevant* to the problem in hand, and that base rates may or may not meet this requirement, depending on the particulars of the situation. This seems to me obviously true, and easily illustrated. Consider the following situation. Kay is a bird-watcher, and it has been established by experiment that she distinguishes ruby-throated hummingbirds from Anna's hummingbirds with an accuracy of about 80%, and is as likely to make the one kind of misidentification as the other. Suppose it is known that ruby-throated hummingbirds are about 19 times as abundant as Anna's hummingbirds (I do not know that to be the case, but made up the ratio for the sake of the illustration). Kay has reported seeing an Anna's hummingbird; what is the probability that she really saw a ruby-throated hummingbird?

The problem is structurally identical to the disease diagnosis problem discussed earlier, and if one uses the 19-to-1 base-rate ratio as the best indication of the prior probabilities of spotting each kind of bird, one gets the same answer: .83. But does it make sense to use the 19-to-1 base rate ratio as the best indication of the prior probabilities of spotting each kind of bird? As it happens, ruby-throated hummingbirds are common in the eastern United States and rare in the west, whereas Anna's hummingbirds are found primarily along the southwest coast and rarely, if ever, in the east. *If* one knew this, *and if* one knew that Kay was in California when she saw what she thought was an Anna's hummingbird, one clearly would not fix the prior probabilities by the 19-to-1 base rate ratio. This base rate would be seen to be irrelevant to the question because it does not reflect the probabilities of encountering birds of the specified types in the particular locale in which the sighting occurred.

Now suppose that one knows (a) that ruby-throated hummingbirds outnumber Anna's hummingbirds, in general, by 19 to 1, and (b) that the first species is found only in the east and the second only in the west, but one has no hint of where Kay was when she saw whatever she saw. What should one use as the prior probabilities in this case? One might argue that the 19-to-1 ratio is the *only* relevant knowledge one has, and that, in the absence of any mo^r^e case-specific information, it should be used on the grounds that it represents the best estimate one has of the ratio of the probabilities of random sightings of the two types of birds. Or one might take the position that if Kay was in the east, what she saw was almost certainly a ruby-throated hummingbird, whereas if she was in the west, it was almost certainly an Anna's hummingbird, and assuming one has no basis for believing the one possibility to be more likely than other, one should take the prior probabilities to be .5. Or, one might convince oneself that the most reasonable thing to do is to set the priors somewhere between these extremes, giving some weight to the fact that ruby-throated hummingbirds outnumber Anna's hummingbirds in the general population, but not as much as one would if it could be assumed that the two species were uniformly distributed over the same territory. In short, I believe there is room for some judgment here and that the argument that there is one and only one correct way to think about this would be a difficult one to make.

If I understand Cohen's position, it is that some base rates are more compellingly appropriate than others and that each case must be judged on its merits. Base-rate statistics are more compelling, for example, when they are about "causally relevant" features of the situation, which is to say when clear causal connections can be made between membership in a class (counts of which define base rates) and events the posterior probability of which is in question. To fail to make such judgments, in Cohen's view, is to fail to recognize that all evidence does not carry the same weight and that, therefore, it should not all be given the same importance. Cohen makes the further claim that some evidence can even be "weight reducing" (as opposed to "weight increasing") in the sense of leading one to use base rates that, upon reflection, are deemed inappropriate.

I agree with this perspective in principle, but recognize that it allows for a larger element of subjectivism in Bayesian reasoning than many people will find appealing. It focuses on the importance of judgments of relevance; it puts the burden on the individual to decide how much weight to give to any bit of evidence that may pertain to a decision or estimate that is to be made. And the rules for making such judgments are not entirely clear. Intuitions, as evidenced by the range of opinions expressed by the commentaries on L. J. Cohen's (1981) article, differ considerably. Moreover, it seems likely that they will continue to differ; it is important to recognize that inasmuch as the probabilities in question are not objectively determinable in most cases, who is right is, and will remain, a matter of opinion. It is interesting, and a little sobering, to note how cocksure many of the experts on different sides of the debate are that they are right.

Returning to the disease diagnosis problem, I find it hard to understand Cohen's argument that, in the absence of a reason for suspecting a greater disposition to one disease or the other in a particular case, one should assume oneself equally disposed to both. My intuition tells me that, in the absence of any more specific information relevant to my specific case, the knowledge that one disease is 19 times more likely than the other in the general population should be taken as evidence that is relevant to my own probable susceptibility. This I take to be the essence also of the position of Blackburn (1981), Krantz (1981), Mackie (1981), Sternberg (1981), and Zabell (1981) among the commentators on L. J. Cohen's (1981) article. I agree with Cohen that general-population statistics provide a less-than-ideal basis for estimating prior probabilities of disease for a given individual, and that it would be much better to be able to use statistics that are more specific to a subset of the population that resembles the individual in question with respect to health-relevant characteristics, but if the latter statistics are not available, it seems to me wrong to ignore the former. They are admittedly "lightweight" evidence, and I might be inclined to modify the 19-to-1 ratio in one or the other direction to the extent that I believed myself to differ from the average person in ways that might be expected to impact susceptibility, but I would not ignore it.

It seems clear that it is important to identify the correct population to use for base rates, and easy to get this wrong. The incidence of some disease in the general population may be, say, 1 in 10,000, but that would be an appropriate ratio to use to evaluate the effectiveness of a diagnostic screening test only on the assumption that testing is done on a random subset of the entire population. If the testing is done on a nonrandom subset—for example, of high-risk or self-selected individuals—a different, and larger, ratio would be appropriate.

DIRECTIONALITY AND WEIGHT OF EVIDENCE

Especially in the context of probabilistic reasoning, theorists have distinguished between what might be called the directionality of evidence and the weight of evidence. Peirce (1932), for example, stressed the importance of both. As was noted in the foregoing discussion regarding the use of base rates as estimates of prior probabilities, some contemporary theorists have argued that in using base rates uncritically, without taking account of the strength of the causal connection between membership in a base-rate class and the events of interest, one, in effect, gives equal weight to evidence that should count for little and to evidence that should count for much.

The distinction is an intuitively appealing one. Given that the evidence in hand favors a particular conclusion, it seems right that the evidence should count for more if it is seen as being highly relevant to the question and very reliable than if it is seen as being only marginally relevant or of dubious reliability. At least our confidence in any conclusion drawn should, it would seem, be greater in the former case than in the latter.

Something of this distinction is captured in the U.S. Army's prescription for evaluating incoming tactical intelligence reports ("spot reports") independently as to probable accuracy of contents and reliability of source (Combat Intelligence Field Manual, FM30-5). The effectiveness of the procedure and the ability of personnel to make independent judgments of accuracy and reliability have been questioned (Baker, McKendry, & Mace, 1968; Samet, 1975a, 1975b), but the desirability of having some indication of how seriously a report should be taken seems clear.

Is this intuition reflected in Bayes's rule? Does Bayes's rule take into account weight of evidence? From one point of view the answer appears to be no. A given conditional probability will have precisely the same effect in updating a prior probability irrespective of the amount of evidence on which the prior is based. Returning to the problem of deciding from which of three urns balls are being drawn, suppose the posterior probabilities associated with the three possibilities are p_1, p_2, and p_3 after n draws. A white ball on the following draw will modify the values of these probabilities by the same amounts, independently of the size of n.

There is another way to look at the situation however. Consider again a Bayesian trying to decide from which of several equally probable urns, containing black and white balls in different ratios, balls are being drawn. Any sequence of several drawings from a real urn containing balls in one of the hypothesized ratios is very likely to have moved the posterior probability of the correct hypothesis closer to 1 and those of the incorrect hypotheses closer to 0. The weight of the evidence obtained from n drawings is reflected in the distribution of posterior probabilities after that number of drawings has been made in the sense that, for a given distribution of prior probabilities, the larger that n is, the closer to 1 the posterior probability of the correct hypothesis is likely to be. It is true that a particular distribution of posteriors will be changed by the same amount by the drawing of a ball of a given color, irrespective of how that distribution was derived, but one is unlikely to *have* the same distribution after n draws as one had after, say, $n/2$ draws.

When one says that according to a Bayesian analysis the distribution of probabilities over a set of hypotheses is thus and so, one has said all that is necessary to continue the analysis; it is not relevant to know the number of observations on which that distribution of probabilities is based. That distribution implicitly reflects the weight of whatever evidence was used to get it to what it is. One who wished to argue that Bayesian analysis does not take weight of evidence into account can point out that a distribution of probabilities that was arbitrarily assigned, or the result of pure whim, has exactly the same standing in the subsequent application of Bayes' rule as one that is the result of several previous applications of it. This is true, but the position also can be defended that when evidence *is* applied, its weight is represented in the distribution that its application effects.

LIMITATIONS OF BAYESIAN REASONING

> All good Bayesian statisticians reserve a little of probability for the possibility that their model is wrong. (DeGroot, 1982, p. 337)

The Bayesian approach to reasoning about probabilistic situations has much to recommend it, and it is not surprising that it has received so much attention from both philosophers and psychologists, or that it has been applied to practical advantage in many contexts. In part just because the approach is so attractive and useful in many ways, it is important to be aware of its limitations, of which there are several.

In a discussion of these limitations a distinction should be made between limitations inherent to the Bayesian approach and other limitations that may result from the ways in which people apply it. With respect to the second type of limitation, Fischhoff and Beyth-Marom (1983) identify a variety of potential sources of bias. According to their analyses, biases can affect about any aspect of the Bayesian approach, of which they identify seven: hypothesis formation, assessing component probabilities, assessing prior odds, assessing likelihood ratios, aggregating data, information search, and action selection.

This list suggests a fairly broad connotation of Bayesian reasoning including, as it does, such activities as information search and action selection, which, strictly speaking, are not distinguishing aspects of the Bayesian approach. Indeed, the only aspect of decision making that uses Bayes's theorem directly is hypothesis evaluation—specifically, the updating of posterior probabilities. Many of the biases that affect performance in a Bayesian decision-making context are not unique to this context, but may be found also when decision making is done without reference to Bayes's rule (Krischer, 1980; Politser, 1981). Here I will deal only with the question of limitations that are inherent to a Bayesian approach.

Conditions of Applicability

The main risk associated with Bayesian reasoning, as I see it, is the risk of expecting more of it than it can deliver. Bayes's rule was advanced to answer a quite specific question: how to revise probabilities associated with existing hypotheses in the light of new data relevant to the truth or falsity of those hypotheses. Given the conditions under which the rule is supposed to work, it appears to do a very good job. Those conditions are restrictive, however, and often are not realized in real-life situations.

The requirement of a well-structured problem in which one has an exhaustive and mutually exclusive set of hypotheses about the possible states of the world is perhaps the most obvious restriction. The assumption that one and only one of those hypotheses is true does not allow for the possibility that none of them is true, or that several of them are true, or partially so.

Inasmuch as the rule specifies only how to revise existing probability estimates, one must begin by specifying prior probabilities if they have not already been specified. Prior probabilities for one problem may, of course, be posterior probabilities of a preceding problem, and this process may be iterated many times. But, as Rozeboom (1997) points out:

> On pain of infinite regress, this process must have a non-Bayesian origin. And there is no coherent credibility distribution assigning certainty to evidence *e* that cannot be reached by conditionalization on *e* from some other coherent credibility distribution. So if a theory of rational belief can advocate no normative standards beyond Bayesian ideals, neither has it any grounds for preferring one coherent allocation of posterior belief strengths compatible with our accumulated data to another. (p. 345)

Rozeboom is quick to point out that this does not defeat the Bayesian perspective, but it does show the insufficiency of that perspective, in the absence of additional principles, to justify belief acquisition.

If one has no rational basis on which to estimate prior probabilities, one must do so arbitrarily. In the absence of any reason for believing otherwise, one might assign equal probability to all the hypotheses under consideration, according to the principle of insufficient reason, as first expressed by Jacob Bernoulli, but, as Luce and Raiffa (1957) have pointed out, indiscriminate use of this principle has led to many nonsensical results. Jaynes (1968) proposed use of the principle of maximum entropy, according to which the assignment is done in such a way as to maximize entropy without contradicting the available prior information. One must decide too, how many hypotheses one should consider and what they should be; Bayes's rule provides no guidance for this task.

Extra-Bayesian Requirements

The modification of beliefs or opinions often involves changing hypotheses rather than revising probabilities associated with existing hypotheses. Bayes's rule prescribes how one is to function within a given problem structure—how one is to apply new data to a given hypothesis set. The ability and willingness to discard an existing structure in favor of one that better fits the facts have been seen by some as defining characteristics of original thinking (Mackworth, 1965; Polanyi, 1963). Bayes's rule has no prescription for modifying problem structure; it makes no provision for changing one's hypothesis set. One is not precluded from deleting hypotheses from the set in hand or from adding hypotheses to it, but when any such change is made, the distribution of probabilities over the set must be modified to take account of the deleted or added hypotheses, and Bayes's rule does not prescribe how this is to be done.

Bayes's rule does not permit revision of a belief to which one attaches certainty—a belief that has a prior probability of 1 or 0. This may be all right if 1 and 0 mean certainty in an objective or definitional sense, but not when they represent subjective certainty. One can be certain but wrong. This aspect of Bayes's rule complicates its application to juror decision making. If one understands the "presumption of innocence" rule to mean that before hearing evidence, one should consider the probability that one is innocent to be 1, then no matter how incriminating the evidence seems to be or how much of it there is, application and reapplication of Bayes's rule to it can produce only a posterior probability of guilt of 0.

There is a certain degree of arbitrariness in the Bayesian approach as to what constitutes a datum. Typically data are associated with the occurrence of observable events, but not all observable events that could conceivably have some bearing on a hypothesis are necessarily considered. Moreover, the nonoccurrence of an event can also be informative, and this seldom is taken into account in Bayesian analyses.

Bayes's rule does not tell one when to stop collecting data for the purpose of evaluating a hypothesis set. Presumably one should cease collecting such data when the cost of obtaining additional data is greater than the value of the information one can reasonably hope to attain. A variety of criteria for deciding when to stop collecting data have been proposed, but all are based on assumptions and arguments that are independent of Bayes's rule.

Finally, the posterior probabilities produced by Bayes's rule are only as good as the prior and conditional probabilities from which they are computed. It is possible to show that these can be quite good—a reasonable basis for decision making—in situations in which they can be determined objectively, which typically means situations involving sufficiently many occurrences of events of interest that relative frequencies can be taken as indicative of the probabilities in question. But many of the real-life situations to which Bayesian reasoning might be applied involve one-of-a-kind or very-few-of-a-kind events, and this precludes verification of assumed probabilities by relative frequency counts; application of Bayes's rule in such cases requires a leap of faith, or at least the assumption that because it works well in situations in which probabilities can be equated with relative frequencies it will work equally well in situations in which they cannot.

ADVANTAGES OF BAYESIAN REASONING

By highlighting limitations of Bayesian reasoning, I do not mean to suggest that the approach has no merits. To the contrary, I believe it to be very useful when appropriately applied. I believe too that its usefulness is enhanced when its limitations are recognized.

The Discipline of Explicit Problem Representation

A major advantage of the use of Bayes's rule to evaluate hypotheses is the requirement it imposes that all hypotheses of interest be explicitly identified and that the application of data to the evaluation of any one of them must take into account the implications of those data to competing hypotheses. Bayesian analysis constitutes a method of applying data to the task of judging the relative tenability of each of a set of hypotheses. It requires that more than one hypothesis be under consideration and has nothing to offer to the individual who has only a single hypothesis in mind. This might be seen as a problem inasmuch as it forces one who would use the approach to come with more than one hypothesis to account for a situation of interest, but this is really an advantage just because it does force one to consider possible alternatives to what may be a favored point of view.

At the very least one must consider not only a hypothesis of interest, H, but the complementary hypothesis, $\sim H$, as well. This is an important discipline. Much evidence indicates that a common failing in reasoning is to evaluate the credibility of a hypothesis by considering the probability of observing some specific data if that hypothesis were true, $p(D \mid H)$, without considering the probability of observing the same data if that hypothesis were false, $p(D \mid \sim H)$ (Nickerson, 1998). The Bayesian prescription precludes this by requiring that an exhaustive set of hypotheses be considered and that evaluation involve the application of the same data to the entire set at the same time.

Relative Insensitivity to Initial Probabilities

Another advantage is the relative insensitivity of the distribution of posterior probabilities to the distribution of prior probabilities, when the rule is applied iteratively over a series of observations. This advantage applies, of course, only when a series of observations is possible; however, the series need not be very long.

I noted as a limitation of Bayes's rule the fact that it does not permit revision of beliefs to which one attaches certainty. Conversely, given prior probabilities other than 0 or 1, the posterior probabilities can never attain 0 or 1. They can get arbitrarily close but never all the way. The latter feature might be seen as a limitation by some, but it can also be seen as a benefit; given that one starts with prior probabilities that are neither 0 nor 1, it prevents one from getting to a point of no return, but leaves open the possibility of a change of mind, no matter how confident one may have become that a particular belief is true.

Cumulative Use of Evidence in Hypothesis Testing

Many critics of conventional statistical significance testing argue that the use of Bayes's rule is a preferred approach to the evaluation of hypotheses (Edwards, Lindman, & L. J. Savage, 1963; Gelman, Carlin, Stern, & Rubin, 1995; Greenwald, 1975; Lindley, 1984; Rindskopf, 1997; Rouanet, 1996; Rozeboom, 1960; Rubin, 1978). One argument in favor of this view is that the use of Bayes's rule permits evidence to be applied simultaneously to a set of hypotheses of interest, and not just to a null hypothesis and its complement. And it yields an update of the probabilities of all the hypotheses, not just a binary decision with respect to a null hypothesis. Also, unlike conventional statistical significance testing, Bayes's rule represents a means of cumulating the effects of evidence on the hypotheses of interest from a series of studies conducted over time. In theory at least, it allows for the posterior probabilities of one set of studies to be the priors for a subsequent one, and the process can be iterated indefinitely. To date, this advantage seems not to have had a large effect on experimental design.

SUMMARY

Bayes's rule or theorem is a prescription for a quantitative form of inductive reasoning, or reasoning from effects to causes. It is sometimes known as the inverse probability theorem. More specifically, it is an equation for revising probabilities to take account of new data. Although Bayes's rule has the limited role indicated, the term "Bayesian decision making" is often intended to encompass all aspects of an approach to decision making that includes application of the rule as a part.

The legitimacy of the rule has been questioned often and its popularity among statisticians and other potential users has waxed and waned since its formulation in middle of the 18th century, but it has many advocates and its popularity appears to have been on the ascendancy over the recent past. Its application is less controversial when the probabilities that are used in the equation are based on objective relative frequencies than when they reflect subjective estimates.

Bayes's rule has both strengths and limitations. It can be used to good effect, but its misapplication can lead to nonsensical results. It should not be applied in a mechanistic fashion. As is true of any approach to statistical reasoning or decision making, its effective use requires a generous dose of good judgment.

CHAPTER

5

Some Instructive Problems

Probability lends itself to a variety of misinterpretations and misunderstandings. Few people would have difficulty understanding, at least in a practical sense, such statements as "The probability that a toss of a fair die will result in a 5 is 1/6," and "The probability that the children in a five-child family are all boys is 1/32." It is easy to describe fairly simple probabilistic situations, however, that can confuse, at least momentarily, even people with a considerable degree of mathematical sophistication. Usually careful reflection on the situation suffices to clarify it. The first few problems described in what follows are of this type.

There are also probabilistic reasoning problems on which people who are very well tutored in statistics and probability theory have been known to disagree. The last several problems described herein are representative of those that may be in this category. It is my belief that the difficulties usually arise because of a lack of sufficient clarity and precision in the use of language and can be resolved by a careful consideration of what could be meant by the problem statement and how it is interpreted by each of the parties who disagree. Several of these problems are discussed also by Falk (1993).

A SMALL WAGER

Imagine the following gamble. A coin is to be tossed repeatedly until one of the two following sequences occurs: (A) head, head or (B) tail, head. Depending on which of these sequences occurs first either you or your opponent wins the game. You are given your choice of which sequence is to be yours. Should you have a preference for one over the other? In fact you should. The odds are 3 to 1 in favor of sequence B and this is so despite the fact that in a long series of tosses each of these two sequences is equally likely.

Why should one prefer sequence B? Consider the four possible outcomes of the first two tosses: HH, HT, TH, and TT. If the first of these occurs the game is over and the player with sequence A wins. The game is also over if the third event occurs and in this case the player with sequence B wins. Assuming a fair coin, these events are equally likely and each occurs with a probability 1/4. If either Events 2 or 4 occurs, the game proceeds, since neither of these is a winning combination. But in both cases, it is now impossible for the combination HH to occur before TH, because in each case the occurrence of the first head will necessarily follow a tail and consequently terminate the game with a win for the player who has sequence B.

This gamble illustrates how easy it is to overlook critical aspects of a probabilistic situation and to jump to erroneous conclusions by making inferences from simple situations to slightly more complicated ones without noticing the complication. What is it that makes the illustrative gamble so easy to misperceive? When one tosses a fair coin twice, the probability of getting a tail followed by a head is precisely the same as getting two heads in a row. Now imagine the following situation. A coin is to be tossed repeatedly and a record kept of the sequence of heads and tails that is produced—TTHTHHHTHTT ... Suppose that this time we decide to enter this sequence at a random point, by picking a number from a hat, say, and check the next two outcomes following that point. If either of our two combinations is found, the game is over. If neither is found, we enter the sequence at another point, determined by the drawing of another number from the hat, and again check the next two outcomes. And so on. In this case, one should have no preference between HH and TH, because they are equally likely to be found, no matter how long the game goes on.

The situation described first is analogous to entering the sequence at a randomly determined point, checking the first and second outcomes following that point and, if neither of the game-terminating pairs is found, moving on to check the second and third outcomes, then the third and fourth, and so on. The second pair of outcomes in this case is not independent of the first, but has one member in common with it. And given the rules of the game, that member is necessarily a T, so inspection of the second and third outcomes can only yield TH that will terminate the game with a win for B, or TT that will mean a repetition of the cycle.

This type of difference and the ease with which it is overlooked is the basis of a wagering scam that can take a variety of forms. In one of them you are to make a wager with an opponent as to which of two three-outcome sequences will occur first in a series of coin tosses, and your opponent, being polite, gives you the opportunity of choosing first. As in the two-toss example just considered, some sequences have a better chance than others of occurring first in an extended set of tosses, and if your opponent is aware of the differences and you are not, you are likely to lose.

Suppose Jack and Jill are playing this game and Jack, on the assumption that all possible sequences are equally likely, bets on HHH. Jill then, being smarter than Jack, picks THH. These sequences are equally likely to occur on the first three tosses, and each has probability 1/8 of doing so. However, if the game goes beyond three tosses, Jill is sure to win. The only way that Jack can win on the first three tosses is for there to fail to be at least one T among them, and given the occurrence of a T, it becomes impossible for Jack *ever* to win, because THH must then occur before HHH can. So, THH will win over HHH seven times in eight; or equivalently, the odds are seven to one in favor of THH.

It is easy to see the advantage of THH over HHH. In the cases of some of the other sequence pairs the relative advantages of one over the other is not so obvious, but invariably one of the pairs does have an advantage. Here I will trace out a hypothetical example for just a few steps to make the idea plausible. Suppose, for example, that Jack had chosen HTH and Jill HHT. Which of them, if either, has the advantage? One way to represent the situation is with a tree, each level of which represents a toss of the coin. Figure 5.1 shows all 32 possible sequences of five tosses

Figure 5.2 is drawn so as to highlight the nine ways in which the game can be terminated within five tosses, two after three tosses, three after four, and four after five. Inasmuch as the probability of arriving at a particular point in this tree as a consequence of n tosses is $1/2^n$, the probability that this game would terminate within five tosses is $(2 \times 1/2^3) + (3 \times 1/2^4) + (4 \times 1/2^5) = 9/16$. The probability that HTH would win within the first five tosses is 1/8 + 1/16 + 1/32

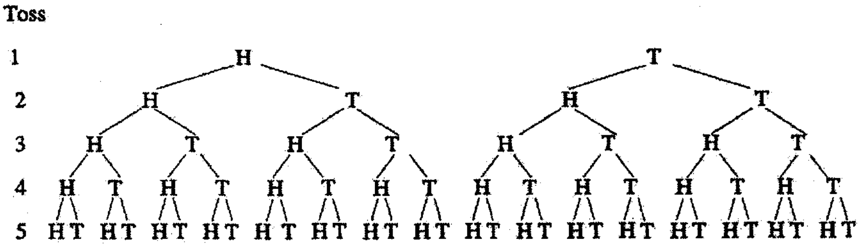

Toss
1
2
3
4
5
H
T
H T H T
H T H T H T H T
H T H T H T H T H T H T H T H T
HT HT HT HT HT HT HT HT HT HT HT HT HT HT HT HT

FIG. 5.1. All possible outcomes of five tosses of a coin.

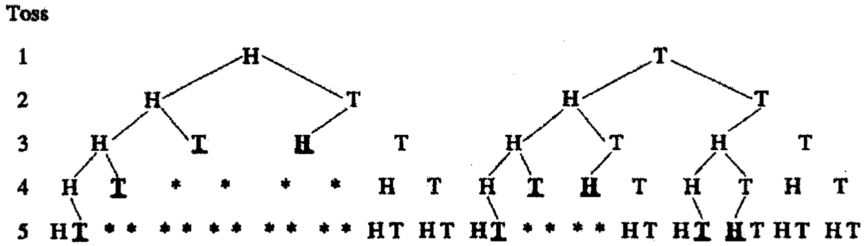

Toss
1 H T
2 H T H T
3 H T H T H T H T
4 H T * * * * H T H T H T H T H T
5 HT ** ** ** ** ** ** HT HT HT ** ** HT HT HT HT HT

FIG. 5.2. All possible ways in which game can be terminated by the occurrence of HHT or HTH within five tosses.

= 7/32, and the probability that HHT would win in the same number of tosses is 1/8 + 2/16 + 3/32 = 11/32; in other words, HHT would have had the advantage, the odds in its favor being 11 to 7, or 1.57 to 1.00. It is obvious that the probability that there will be a winner increases with the number of tosses that are allowed; the reader may wish to verify that the extent to which the odds favor HHT get better the longer the game is allowed to go on. The odds favoring HHT are 1.69 to 1.00 if the game is allowed to go to 6 tosses and 1.92 to 1.00 if it is allowed to go to 10.

This example perhaps suffices to make it intuitively clear that two triplets can have the same probability of occuring in three tosses and yet one of them be much more likely than the other to occur first in an extended series of tosses. What is probably less intuitively clear, but no less true, is the fact, first noted by Penney (1969), that given a specified triplet, there is *always* another triplet that has an advantage over it in an extended series of tosses. In other words, if Jill understands the situation completely, she can always pick a triplet that is a better bet than the one Jack picked, no matter what triplet Jack picked. Hombas (1997) has presented a detailed analysis of the situation. Table 5.1 gives the results of his analysis and shows for each possible combination of triplets chosen by two players, A and B, the probability that the triplet chosen by the second player to choose, B, will occur before the one chosen by the first player to choose.

Player B's optimal strategy is to pick the triplet corresponding to the row that contains the largest probability in the column corresponding to the triplet picked by A. Suppose, for example, that A chooses TTH. Then B should choose HTT, in which case the latter's chances of winning would be 3/4, or odds of 3-to-1. If A chose either HHH or TTT, B could get his probability of winning to 7/8, or odds to 7-to-1, by choosing THH in the former case and HTT in the latter. A can limit B's potential advantage to odds of 2-to-1 by choosing HTH, HTT, THH or THT, but no matter what A chooses, B can ensure favor-

TABLE 5.1

Probabilities That the Triplet Chosen Second by B Will Occur Before the Triplet Chosen First by A

	A							
B	*HHH*	*HHT*	*HTH*	*HTT*	*THH*	*THT*	*TTH*	*TTT*
HHH	*	1/2	2/5	1/8	5/12	3/10	1/2	1/2
HHT	1/2	*	2/3	2/3	1/4	5/8	1/2	7/10
HTH	3/5	1/3	*	1/2	1/2	1/2	3/8	7/12
HTT	3/5	1/3	1/2	*	1/2	1/2	3/4	7/8
THH	7/8	3/4	1/2	1/2	*	1/2	1/3	3/5
THT	7/12	3/8	1/2	1/2	1/2	*	1/3	3/5
TTH	7/10	1/2	5/8	1/4	2/3	2/3	*	1/2
TTT	1/2	3/10	5/12	1/8	2/5	2/5	1/2	*

Note. After Hombas (1997).

able odds of at least 2-to-1. The solution represented in Table 5.1 illustrates a type of intransitivity that can arise among probabilistic relationships. Note that HHT beats HTT, which beats TTH, which beats THH, which beats HHT, which is where we started. Other intransitivities involving probabilistic relationships will be noted in chapter 6.

THE KING'S FOLLY

There once was a king in a polygamous land who worried about the fact that men and women were born in about equal numbers. This bothered him because the ideal situation, in his view, was for there to be many more women in the realm than men so that every man could have several wives. As it was, the men who were successful in acquiring several wives were resented by those who were not and this caused a lot of friction and strife. The king's predecessors had solved this problem by periodically sending the men off to war, thus ensuring that females would outnumber males in the realm by a large margin despite the near-equal birth rates of the two genders.

This king had no taste for war, however, and believed there must be a more humane way to accomplish his goal. After thinking about the problem for a long time and consulting his advisers, he decided to issue a decree that any family could have as many children as it wished until it produced its first boy, whereupon it was to cease and desist from further reproduction; having a child after already having had a boy would result in banishment from the land. Furthermore, to motivate families to keep having children as long as they were producing girls, he provided a most generous monetary award upon the birth of every female child. The king reasoned that his decree would guarantee that no family in his realm would have more than one boy, and that his incentive policy would ensure that many families would have more than one girl, and some would have several.

The decree was issued and strictly enforced, and the awards were made without fail. The king observed the results of his action for a few years with great satisfaction. Considering only the children born after the issuance of the decree, one could not find a family in the realm that had more than a single boy, but one could find many families with two, three, four, or more girls. The king was ecstatic to learn that in his realm of a few tens of thousands of people there were a few families that had had as many as 10 girls before having to call it quits because of finally having a boy!

So pleased was the king with how well his policy was working that he decided to quantify the results so they could be reported to the people. He commissioned a census aimed at counting the number of boys and girls in the realm that had been born since the issuance of the decree. To the king's dismay, the count, and the recount, showed the number of male children born after the decree to be about the same as the number of female children born during the same time. Strict compliance with the decree had had no effect on the ratio of male-to-female births in the realm, which had remained one-to-one.

Where had the king gone wrong in his thinking? It was true that an effect of the decree was to guarantee that no family in the realm had more than one boy. And it was the case that there were many families with more than one girl. Where the king had gone wrong was in assuming that these effects were tantamount to a larger number of female births than of male births.

I have not tried to determine in any very formal way whether most people would share the king's surprise at the ineffectiveness of his attempt to increase the ratio of women to men in his kingdom. The reactions of a few people to whom I have told the story make me suspect that many people would do so. If this conjecture is correct, a question that arises is whether there is an insight that can make the situation clear.

Assume, for the sake of simplicity, that every family in the realm has as many children as possible without violating the decree, and that the probability that any random child will be a boy is exactly 1/2. One insight that may help make the outcome plausible is that whereas *every* family will have one boy, one half of all families—those who produce a boy as the first child—will have *no* girls. Specifically, the distribution of girls over families in the realm will be as follows: One half of all families will have no girls, one half will have one or more girls, one fourth will have two or more girls, one eighth will have three or more, and, in general $(1/2^n)$th will have n or more. As it happens, the number of boys produced by having one boy in every family turns out to be about the same as the number of girls produced by this distribution.

Mathematically, the result follows from the fact that

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \cong 1.$$

So, X families in the realm would be expected to produce X boys and

$$\sum_{n=1}^{\infty} \frac{X}{2^n} \cong X,$$

girls.

The process may be clarified with a concrete illustration. Imagine 1,000 families, each producing a child, yielding a total of 500 girls and 500 boys. The families that produced boys can have no more children, but each of those that produced a girl the first time go on to have a second child; 250 of these produce girls and 250 boys. We now have a total of 750 girls and 750 boys. The 250 families that had two girls go on to produce a third child, and so on. The king's decree clearly had an effect on family structure in the realm. It ensured that no family would have more than one boy, and that only parents who produced strings of girls had sizable families. It had no effect, however, on the ratio of girls and boys in the realm.

The reader who has difficulty with this story may wish to do an analogous experiment. Consider a sequence of coin tosses to be terminated with the first occurrence of a head. Thus H is one such sequence, TH is another, TTH another, and so on. Now toss a coin many times, keeping track of the sequences that occur. After a few thousand tosses, you will find that about half of your sequences are one toss long (H), about one quarter are two tosses long (TH), about one eighth three tosses long (TTH), and, in general, about $1/2^n$ are n tosses long. Although many sequences have several Ts, none has more than one H. The total number of Hs and Ts in all the sequences combined, however, will be about equal. This should be intuitively obvious from the fact that in any large number of tosses of a (fair) coin, the number of heads should approximately equal the number of tails, and in conducting this experiment all one does is toss a coin a large number of times. Perhaps it is clear from this example that there is no "contingent-termination" method that will be generally effective in determining the percentage of outcomes of a given type (say tosses of heads) of a random process.

A SUBTLE DISTINCTION WITH UNSUBTLE IMPLICATIONS

In a certain town, the average family has three children; it follows that the average child in this town has two siblings. If you did not object to this inference, you are probably in good company. It sounds right, but it is not. The number of siblings that the average child in this town has is likely to be much greater than two; it would be two only in the unlikely event that *every* family has three (the average number) of children, in which case *every* child has two siblings. If there is any variability in family size, the average child has more siblings than does a child from an average-size family.

J. J. Jenkins and Tuten (1992) discuss this situation and note that the fact that must be grasped to see through the puzzle is that the average child does not come from the average family. Imagine that there are 10 families in this town, 5 of which have one child each and 5 of which have five each. In the aggregate the 10 families have 30 children, which gives us an average of 3 per family (although in this example, there is no actual family that has the same number of children as the "average family"). If each family had the average number of children, three, each child would have two siblings. But each of the 5 children in the 1-child families has no siblings, whereas each of the 25 children in the 5-children families has four siblings, so the number of siblings of the average child is $[(5 \times 0) + (25 \times 4)]/30 = 3.33$.

One might object that language has been used a little loosely in this discussion. What, after all, is an "average family" or an "average child"? Does it make sense to speak of an "average family" in a situation, like our imaginary one, in which *no* family in the population of interest is "average?" And one could argue that "average child" should be taken to mean child from an average family, although that is not the way this term was used in our example.

To be more precise, we could reword the original assertion as follows. In a certain town, the average number of children per family is three; it follows that the average number of siblings per child in this town is two. This inference is, of course, incorrect, and having just thought through the problem as originally worded, we are likely to see that straightaway. Whether people will generally be more apt to see the problem with the inference when stated in the second, more precise, way than when stated in terms of average families and average children will be left to the reader to check. My suspicion is that many people will see no problem with the inference in either form, until it is pointed out to them, but the second wording does have the advantage over the first that its use of the terms "per family" and "per child" makes explicit the fact that number that is used as the denominator to calculate the "average" differs in the two cases.

Failure to make this type of distinction can lead to the drawing of unwarranted conclusions or present the opportunity for deception. Consider, for example, the question of average class size in a school district, a question of considerable practical interest to parents and school administrators in many towns. One is inclined to assume that the smaller the average class size, the better. The reason, or at least *a* reason, for this assumption is the belief that a child is likely to get more personal attention in a small class than in a large one. But it is important to recognize that *average number of pupils per class* is not the same as *average number of classmates per pupil.* Consider the two following situations. School District A has 20 classrooms, each housing a class of 20 students, whereas District B has 5 classrooms with 5 students, 5 with 10, 5 with 25 and 5 with 40. Both districts have an average of 20 pupils per class; however, the average number of classmates per pupil is 19 in District A and 33.4 in District B.

J. J. Jenkins and Tuten (1992) point out that the problem represented by the average family–average child puzzle has important implications for the interpretation of descriptive statistics. By way of illustration, they cite the following example noted by D. S. Smith (1979): "The mean population of the 433 Midwestern counties in 1900 was 36,853. The average person in the Midwest in 1900, however, had another 284,821 persons living in his county; a sample of Midwestern counties would produce a poor sample of Midwesterners" (p. 85). Jenkins and Tuten note too that the distinction has implications for sampling: If one wants to select a representative sample of families, one generally will not be assured of getting one by randomly sampling children. Moreover, they argue, the problem is a very general one: "The average churchgoer does not attend the average-size church; the average person does not live in the average-size household; the average club member does not belong to the average-size club; the average rat pup is not from the average-size litter; ... In the most general sense, we must be aware that the average individual classified by some system will not be in the average-size class of that system" (p. 524).

KNOWING WHEN TO QUIT

> It is an indubitable result of the theory of probabilities that every gambler, if he continues long enough, must ultimately be ruined. (Peirce, 1956, p. 1337)

To illustrate the futility of gambling indefinitely, Peirce used the example of the "Martingale," in which the player doubles her bet following every loss. Starting with a bet of $1, if she loses three bets in a row and wins the fourth, she loses 1 + 2 + 4 or $7 and wins $8. Inasmuch as

$$\sum_{i=0}^{n-1} 2^i + 1 = 2^n ,$$

the player is assured in this game to make up any string of consecutive losses with a single win, if she has enough money to last through the losing string. But sooner or later she will encounter a sufficiently large loss that she will not have enough money to cover it and thus will be ruined. And it does not matter how much money she wins before the ruin, because just as she can recover all losses with a single win, so can she lose all winnings with a single loss.

Peirce's illustration is not an argument against gambling per se, but against continuous gambling with an opponent who has a much larger bank than you. It applies to compulsive gambling with a well-financed gaming establishment, unless you are rich enough to be the owner. It does not apply to the gambler who is careful to gamble only against poorer opponents. Keynes 1921/1956) makes this distinction explicit in a discussion of the importance of an adequate bankroll in gambling, in which he notes that "the poorer a gambler is, relatively to his opponent, the more likely he is to be ruined" (p. 1370). The moral that Keynes extracts from this observation is "that poor men should not gamble and that millionaires should do nothing else" (p. 1371).

The gambler who is free to specify the size of the wager on every bet and to terminate the betting whenever he likes is—barring an urge to self-destruction—almost certain to win. Consider again the Martingale, in which every bet is double the amount of the preceding one. Table 5.2 shows the results of a sequence of 30 (actual) tosses of a coin, given an initial bet of $1, a doubling of the stake on each toss, and head (H) being defined as a win. The fourth column shows the gambler's net gain or loss as the game proceeds. Note that the net is always positive following the toss of a head and always negative following the toss of a tail; this is because the absolute value of the net at any given point must be smaller than the stake for the next bet.

To (almost) ensure winning in this situation all one needs to do is decide to stop after tossing a head. One need not stop after the first head, of course, or even after the first head following a run of tails, but one should stop before the stake gets large enough that a loss could break one's bank. That is why I say "almost certain" rather than simply "certain"; there is always the possibility, however remote, of an initial sequence of losses long enough to wipe one out. (There is also the problem of being unable to stop when a win on the next bet would be so desirable even though a loss would be ruinous.) In the successive doubling situation, the length of the sequence that would be required shortens as the game goes on; indeed the stake gets very large very rapidly, so one can easily get wiped out by a single loss if one lets the game go on too long. A single loss is *always* more than large enough to wipe out all previous winnings; each negative number in the table indicates the cumulative loss following the associated bet, net of any winnings along the way.

TABLE 5.2

Results of Hypothetical Sequence of Bets With a Doubling of the Stake on Each Bet

Bet #	*Stake*	*Toss*	*Net*
1	1	H	1
2	2	H	3
3	4	H	7
4	8	T	–1
5	16	H	15
6	32	H	47
7	64	T	–17
8	128	H	111
9	256	T	–145
10	512	H	367
11	1024	H	1391
12	2048	H	3439
13	4096	T	–657
14	8192	H	7535
15	16,384	T	–8849
16	32,768	H	23,919
17	65,536	T	–41,617
18	131,072	H	89,455
19	262,144	H	351,599
20	524,288	T	–172,689
21	1,048,576	H	875,887
22	2,097,152	H	2,973,039
23	4,194,304	T	–1,221,265
24	8,388,608	T	–9,609,873
25	16,777,216	H	7,167,343
26	33,554,432	T	–26,387,089
27	67,108,864	H	40,721,775
28	134,217,728	T	–93,495,953
29	268,435,456	H	174,939,503
30	536,870,912	T	–361,931,409

Here is a rather different problem having to do with deciding when to quit. Imagine that you are a personnel manager of a company and you have the responsibility of hiring a new secretary for one of the company's departments. Suppose there is a pool of *N* candidates from which you can make the selection but that you must proceed according to the following admittedly somewhat unrealistic rule. You can consider as many of the *N* candidates as you wish, in random order, but you must consider them one at a time and make a yes–no decision with respect to each candidate before proceeding to the next one. In other words, you get only one chance to consider a given candidate; after you decide to pass a candidate and go on to the next one, you get no opportunity to reconsider the candidate you passed. Assume further that you have *no* advance information about the makeup of the candidate pool—all the candidates could be terrible, all wonderful, or there could be any conceivable mix—and that you get no information about individuals in it before considering them in turn. Assuming you want to make the best possible choice, when should you stop considering candidates and select one?

It seems fairly clear that, given these constraints, it is not possible to ensure that you will select the best candidate, but is there a rule that will maximize your chances of doing so? Intuition may give conflicting signals here. On the one hand, one might take the position that, inasmuch as the order in which the candidates are to be considered is random, the best candidate is equally likely to be any of them, so it makes no difference where you decide to stop. On the other hand, one might reason that if one stops after considering only a small subset, say 10%, of the candidates, the best choice is likely to be among the 90% not yet considered, and conversely if one considers 90% before selecting, the best candidate is more likely to be among those already bypassed than among those that remain.

There is a strategy that will maximize one's chances of making the best choice (Ferguson, 1989; Gilbert & Mosteller, 1966). What one should do is *pass* the first *M* candidates considered and then select the first candidate thereafter (if there is one) that is better than the best among the first *M*. The trick is to determine what *M* should be; if *M* is too small relative to *N*, the chances are that the selection will be made too soon and the best candidate will be among those never considered, whereas if it is too large, the chances are the best candidate will be among the first *M* considered and passed. In fact the optimal choice of *M* is *N/e*, which is to say that one should pass approximately 37% of the candidates and then pick the first one (if there is one) who is better than any of those passed. It turns out that if this strategy is followed, the probability that one will end up with the best candidate is 1/*e*, or approximately .37; *e* pops up in the strangest places.

The fact that this strategy maximizes one's chances of selecting the best candidate does not necessarily make it a rational strategy from all points of view. One might question the wisdom of holding out for the best candidate under these circumstances. If it should happen that the best candidate is among the 37% passed, which will be the case with probability .37, then one will end up with the last candidate in the pool, and the chances that this candidate will be within the top 10% are only 1 in 10. The joint probability that the best candidate is among the 37% originally passed and that the final candidate (who must be chosen by default) is not among the top 10% is (.37)(.9), or about .33. If the best candidate is *not* among the 37% passed, the probability that that candidate will be selected by the prescribed rule is about .59, so the probability that the best candidate will *not* be selected, given that that candidate remains in the pool after the first 37% have been passed, is about .41 and we cannot be sure that the candidate who *is* selected in these cases will be among the top 10%.

The point is that if one takes as one's initial goal of selecting a candidate who is much better than average, but not necessarily the best—if one attempts to satisfice instead of optimize—one may be able to increase one's chances of not being left with a really poor choice by selecting not necessarily the first candidate who is better than any among the first 37% considered, but the first one who is, say, above the *n*th percentile of that group, letting *n* start with a value close to 100 immediately after 37% of the candidates have been considered and drop gradually as one gets closer to exhausting the pool. For alternative simple rules for solving the secretary selection problem, which require considering far fewer than 37% of the candidate pool to establish the criterion, and that outperform the 37% rule in terms of some desiderata (other than maximizing the probability of getting the very best), see Todd and Miller (1999). This is only one of many illustrations that could be given that challenge the wisdom of trying to optimize.

SIBLING GENDER

Some of the problems just described might be mildly surprising to people with a good understanding of probability theory when they first think about them, but they are unlikely to provoke any lasting debate, because the reasons why a superficial analysis might produce the wrong answer are obvious on reflection. Situations can be described, however, for which people who are familiar with probability theory will disagree as to what the probabilities of specified possibilities are. The following example is one of several discussed by Bar-Hillel and Falk (1982): "Mr. Smith is the father of two. We meet him walking along the street with a young boy whom he proudly introduces as his son. What is the probability that Mr. Smith's other child is also a boy?" (p. 109). Bar-Hillel and Falk report that two professors of mathematics, when given this problem, disagreed as to whether the correct answer is 1/2 or 1/3.

The following problem appeared several times in the "Ask Marilyn" column of *Parade Magazine* (Vos Savant, 1990a, 1990b, 1991):

> Suppose you're on a game show, and you're given a choice of three doors. Behind one door is a car; behind the others, goats. You pick a door—say, No. 1—and the host, who knows what's behind the doors, opens another door—say, No. 3—which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? (Vos Savant, 1991, p. 12)

Publication of this problem, and of the answer proposed by Vos Savant (that it is to your advantage to switch), stimulated a great deal of mail, much of it from people with university- or graduate-level exposure to probability theory. Some responders argued strongly for one position, some for the other. Responses often indicated a great deal of confidence in the positions defended, and sometimes showed disdain for the incompetence of anyone who could hold a different view. Since its first appearance in Vos Savant's column, this problem—sometimes referred to as "the player's dilemma" and sometimes as "Monty's dilemma," after Monty Hall, the longtime host of "Let's Make a Deal"—has been discussed by several writers in the literature on probabilistic thinking (Falk, 1992; Gillman, 1992; J. P. Morgan, Chaganty, Dahiya, & Doviak, 1991a, 1991b; Selvin, 1975a,1975 b; Seymann, 1991; Shaughnessy & Dick, 1991).

Problems like those described by Bar-Hillel and Falk and by Vos Savant illustrate in a particularly compelling way that probabilistic reasoning can be tricky, even for people who are well versed in probability theory. What makes these problems, and many others that could be considered, difficult may have a complex answer. However, I want to argue that a major contributing factor is the fact that statements of problems often are incomplete or ambiguous in the sense that they admit of more than one interpretation, depending on assumptions that the reader may make about the situation described, possibly without realizing he or she is making them, in the process of deriving an answer. I have tried to make this possibility plausible by noting the assumptions that are implicit in the different answers to the two problems just mentioned and to others of a similar nature (Nickerson, 1996a). If this hypothesis is correct, one way to reduce disagreements regarding solutions to probability problems among people who are knowledgable about probability theory is to insist that the assumptions on which probabilities are calculated be made explicit. Two probabilistically knowledgeable persons *working on the same assumptions* should not produce different answers to the same problem.

Bar-Hillel and Falk's (1982) analysis of the sibling gender problem points out the relevance of how the information that Smith has at least one son was obtained and notes the need to make an assumption about the probability of meeting Smith on the street with a boy if he happened to have one boy and one girl. If we assume that if Smith has a boy and a girl so that we are equally likely to find him in the company of either—that the child we see with Smith was randomly selected from the two he has—the correct answer is 1/2. In contrast, if we believed him to be sufficiently partial to boys that he would walk only with a son if he has a boy and a girl, then the correct answer is 1/3. On the other hand if we believe him to be partial to girls and willing to walk with a son only if he has no daughter, the correct answer is 1 (Nickerson, 1996, Table 1).

Bar-Hillel and Falk (1982) argue that in the absence of evidence to the contrary, it seems natural to assume that a father of a son and daughter would be as likely to be seen with one as with the other. This seems right to me, but the point is that it is an assumption and should be recognized as such. The point is illustrated by consideration of another possible answer to the question, but one that is justified only if a different assumption is made about the probability that a father of a son and daughter would elect to walk with the son.

Bar-Hillel and Falk (1982) give the following argument as one that might be used in support of the the answer 1/3. Given that Smith is the father of two children, he must be the father of two boys, of two girls, or of one boy and one girl, with probability 1/4, 1/4, and 1/2 respectively. Discovering that at least one of the children is a boy rules out the possibility of two girls and identifies the family as a member of the subset of two-child families that have at least one boy, and we know that about 1/3 of such families have two boys. So, one might conclude, the probability that the other child is a boy is 1/3.

Bar-Hillel and Falk (1982) point out that the conclusion is justified only if another unstated assumption is made, namely that the family not only is a member of the subset of two-child families that have at least one boy but that it is a *randomly selected* member of that subset, which is tantamount to assuming that all members of this subset are equally likely to be represented on the street by a father and son. But this assumption would be reasonable only in a land where fathers who had a son and a daughter would walk only with the son. In a land where fathers with a son and a daughter are as likely to walk with one as with the other, any family with two sons is twice as likely to be represented on the street by a father and son as is any family with only one son.

THREE CARDS

Another problem that illustrates the ease with which people can be confused by conditional probabilities involves three cards, one of which (RR) is red on both sides, one of which (WW) is white on both sides, and one of which (RW) is red on one side and white on the other. Imagine the following scenario. Person A asks person B to shuffle the cards out of sight, pick one at random and report the color of (only) one side. A cannot see the card and neither can you.

B shuffles the cards, picks one and reports "red." A then notes that the card obviously cannot be the white-white one, but that it could be either of the other two. A then offers to bet $10 against 9 that the other side of this card is red. Should you take the bet?

Assuming that B selected *at random* which side of the card to report, the probability that you will lose if you play is 2/3. A has offered you a bet in which his expected gain is (2/3) × $9 – (1/3) × $10 = $2.67 (which, of course, is your expected loss). It is true that in knowing that one side of the card is red eliminates the WW card from consideration, but it does not follow that the card in hand is equally likely to be either of the remaining two cards. The important thing to see is that the RR card is twice as likely to show a red face, if picked, as is the RW card. About two thirds of the college students to whom Bar-Hillel and Falk (1982) gave a version of this problem judged the probability that both sides of the card were red, given one side was shown to be red, to be 1/2.

Here is a way to think of the problem that may help make it less confusing. Imagine that the six sides of the three cards are distinguishable by some way other than their colors. To be specific, let us imagine that the sides are numbered 1 through 6, the sides of the red-red card being numbered 1 and 2, those of the white-white card 3 and 4 and those of the red-white card 5 and 6. If a card is drawn at random and placed on the table so a random side is up, then each *number* is equally likely to be showing following a draw. Let the notation R_1R_2 indicate that the red-red card has been drawn and the side with the number 1 is up. The situation is represented in Table 5.3. It should be clear from the table that in two of the three equiprobable cases in which a red card is seen, the unseen side of the card is also red. So, given that the "up" side of a randomly selected card is red, the probability that the "down" side is also red is 2/3.

The following are two variations on the three-card problem:

A: While blindfolded, you draw all three cards from a hat and place them in a row on the table. When you are allowed to look at the cards you see that, left to right, they are showing red, white, and white. What is the probability that the hidden side of the leftmost card is red?

B: While blindfolded, you draw all three cards from a hat and place them in a row on the table. When you are allowed to look at the cards you see that they are showing two reds and a white. What is the probability that the hidden side of the leftmost red card is red?

One way to solve these problems is to consider all possible outcomes of the drawing of the three cards as represented in Table 5.4. The notation in this case indicates the triplets of sides *showing*. Inasmuch as there are three cards, each

TABLE 5.3

A Card (From Among Three Cards, Red$_1$-Red$_2$, White$_3$-White$_4$, and Red$_5$-White$_6$) Is Chosen at Random and the Color of a Random Side of the Selected Card Is Seen

	Color Seen		
Card	*r*	*w*	*S*
R_1R_2	1/6	0	1/6
R_2R_1	1/6	0	1/6
W_3W_4	0	1/6	1/6
W_4W_3	0	1/6	1/6
R_5W_6	1/6	0	1/6
W_6R_5	0	1/6	1/6
Σ	1/2	1/2	1

of which has two sides, there are 2^3 ways in which the sides can be combined. Each one of these combinations can occur in six different orders, so taking order into account there are 48 equiprobable arrangements of the three cards. All are listed in the table.

To see the answer to Question A, we note that 8 of the 48 possible arrangements of cards show RWW, in that order. In all eight cases, the red face showing is either R_1 or R_2 (never R_5), and both R_1 and R_2 have red on the opposite side, so the answer is that the probability that the hidden side is red is 1. The answer is obvious too if we note that if two white faces are showing, one must be W_3W_4 and the other R_5W_6, from which it follows that the remaining card—the one showing red—must be R_1R_2. This applies independently of the order in which the cards are arranged.

Now consider Question B. Inspection of Table 5.4 will reveal that of the 48 equally probable arrangements, 24 have two reds showing and for 12 of these the leftmost card showing red is red on both sides (R_1 on one side and R_2 on the other), so the answer to the question is 12/24 or 1/2. Again, the answer should be intuitively clear simply from consideration of the fact that if two reds are showing, one of them must be R_1R_2 and the other R_5W_6, and each of these is as likely as the other to be the leftmost of the cards showing red. It should be clear also that the answer would be the same if the question had been, given that when you look at the cards, you see two reds and a white, what is the probability that the hidden side of the rightmost red card is red.

TABLE 5.4

All Possible Arrangements of the Six Sides of Three Cards, Taking Order Into Account

$R_1W_3R_5$	$R_1R_5W_3$	$W_3R_1R_5$	$W_3R_5R_1$	$R_5R_1W_3$	$R_5W_3R_1$
$R_1W_3W_6$	$R_1W_6W_3$	$W_3R_1W_6$	$W_3W_6R_1$	$W_6R_1W_3$	$W_6W_3R_1$
$R_1W_4R_5$	$R_1R_5W_4$	$W_4R_1R_5$	$W_4R_5R_1$	$R_5R_1W_3$	$R_5W_3R_1$
$R_1W_4W_6$	$R_1W_6W_4$	$W_4R_1W_6$	$W_4W_6R_1$	$W_6R_1W_4$	$W_6W_4R_1$
$R_2W_3R_5$	$R_2R_5W_3$	$W_3R_2R_5$	$W_3R_5R_2$	$R_5R_2W_3$	$R_5W_3R_2$
$R_2W_3W_6$	$R_2W_6W_3$	$W_3R_2W_6$	$W_3W_6R_2$	$W_6R_2W_3$	$W_6W_3R_2$
$R_2W_4R_5$	$R_2R_5W_4$	$W_4R_2R_5$	$W_4R_5R_2$	$R_5R_2W_3$	$R_5W_3R_2$
$R_2W_4W_6$	$R_2W_6W_4$	$W_4R_2W_6$	$W_4W_6R_2$	$W_6R_2W_4$	$W_6W_4R_2$

ACES AND KINGS

Another problem that has some features in common with those that have already been discussed prompted a series of articles in *Philosophy of Science* some years ago (A. I. Dale, 1974; Faber, 1976; S. Goldberg, 1976; Rose, 1972). The series is instructive, because it illustrates well how confusing relatively simple probability problems can be even for people who are knowledgeable about probability theory. The statement of the problem that precipitated the series was given by Copi (1968) in an introductory-logic text as follows:

> Remove all cards except aces and kings from a deck, so that only eight cards remain, of which four are aces and four are kings. From this abbreviated deck, deal two cards to a friend. If he looks at his cards and announces (truthfully) that his hand contains an ace, what is the probability that both his cards are aces? If he announces instead that one of his cards is the ace of spades, what is the probability then that both his cards are aces? (These two probabilities are *not* the same!) (p. 433)

Copi (1968) gives 3/11 as the first probability and 3/7 as the second. Rose (1972) accepts 3/7 as the correct probability for the second case and argues that accepting this answer for that case commits one to acceptance of it for the first case as well:

> For if at least one of the cards in the hand is an ace, then either the hand contains A_s (subscript = suit) or it contains A_h or it contains A_d or it contains A_c. But if the hand contains A_s, the probability of two aces is 3/7 If it contains A_h, the probability of two aces is 3/7. If it contains A_d, the probability of two aces is 3/7. And if it

> contains A_c, the probability of two aces is 3/7. Therefore, by constructive dilemma, the probabiltiy that both cards are aces is 3/7, even if no suit has been mentioned. (p. 523)

Rose (1972) claims that:

> The error of those who arrive at a probability of 3/11 when no suit has been mentioned lies in supposing that the various possible *hands* are equipossible, and then noting that of the 22 hands containing at least one ace there are six containing two aces.... But the equipossibles are *not* the hands, for this is not one of those standard questions about probability of a certain *hand;* what is distinctive and crucial about the problem before us is the information that there is *at least one ace*. And *any one* of the 28 *aces* that occur in the above array [Rose had listed the 22 hands containing at least one ace, 6 of which contain two] could provide just as much support for *that* information as could any of the other aces. Thus the equipossibilities are the 28 *aces* in the above array of hands, and not the 22 *hands* themselves; by the same token, the favorable outcomes here are those *aces* that are accompanied by another ace, and not the *hands* in which there are two aces. Therefore, since a total of 12 of the 28 aces are accompanied by another ace, the probability that the hand contains two aces is 12/28 = 3/7, exactly the same as when a suit was named. (p. 523)

Rose's (1972) analysis goes wrong, in my view, in two ways. First, it takes Copi's (1968) statement of the problem as adequate to permit an unequivocal solution, but it is not. We are not told under what conditions the friend to whom the cards are dealt will announce (truthfully) that his hand contains an ace. Rose's analysis assumes that the friend will announce that his hand contains at least one ace whenever it does so. But this is an assumption that need not be true. The friend might announce on the basis of some other rule—he might, for example, be more inclined to announce having a king whenever he had at least one king, in which case he would announce having an ace only when he had two of them. Or he might use some other rule, or no rule at all, deciding whether to announce having an ace or having a king—when he has one of each—on the basis of momentary whim. The point is, in the absence of knowledge of, or an assumption about, his reporting rule, the problem is not solvable. In this regard, the situation is analogous to the coin-tossing and sibling gender problems discussed above. A similar ambiguity applies to the case in which the friend announces having the ace of spades; again, one must know, or make an assumption about, the conditions under which he would announce this in order to know what to make of that announcement.

But now let us make the assumption that the friend's rule is to report having an ace *whenever he has at least one ace*. We might imagine that after dealing the cards, you ask your friend to look at both of them and tell you whether or not at least one of them is an ace. What do we make of Rose's argument that the probability that the hand contains two aces, given that it contains at least one, is 12/28, or 3/7? If we analyze this situation, we get the answer 3/11. There are 28 possible two-card hands that can be dealt from eight cards; 6 will have two aces, 6 two kings, and 16 will have one ace and one king. The entries in Table 5.5 show the relative frequencies with which the friend will report holding an ace.

Each cell in the table represents the probability of the joint occurrence of the hand indicated by the associated row label and the report indicated by the associated column label. Letting $p(\text{AA} \mid \text{a})$ represent the probability that there are two aces, given that there is at least one ace (that the friend reports having an ace), $p(\text{AA\&a})$ the probability of the joint occurrence of two aces in the hand and the report of having an ace, and $p(\text{a})$ the probability of the report of having an ace,

$$p(\text{AA} \mid \text{a}) = p(\text{AA\&a})/p(\text{a}) = (6/28)/(22/28) = 3/11.$$

In other words, the probability of there being two aces, given that there is one ace, is the ratio of the number of hands having two aces (6) to the number of hands having at least one ace (22).

Rose (1972) contends that this ratio is not the appropriate one for this problem. The argument goes as follows. If we consider only the 22 hands that contain at least 1 ace, we see that there are 28 aces in those hands—12 of them in the 6 hands that contain 2 aces each, and 16 in the hands that contain 1 ace and 1 king. The probability of there being two aces, given that there is one ace, is the ratio of the *number of aces in the hands containing two aces* (12) to the *number of aces in the hands containing at least one ace* (28).

Which of these analyses (if either) is correct? Clearly they cannot both be. Imagine that the card game is played a sufficiently large number of times that the relative frequencies of the various possible hands closely approximate

TABLE 5.5

The Rule Is to Report Having an Ace Whenever the Hand Contains at Least One Ace

	Friend's Report		
Hands	*ace*	—	Σ
AA	6/28	0	6/28
AK	16/28	0	16/28
KK	0	6/28	6/28
Σ	22/28	6/28	1

those predicted by theory and the assumption that all possible hands are equiprobable on every deal. The proportion of hands that contain two aces will be about 6/28; the proportion that contain at least one ace will be about 22/28; and the proportion of those that contain at least one ace that contain two aces will be about 6/22, or 3/11.

It is also true that about 12/28, or 3/7, of all the aces that are in hands containing at least one ace (which is to say all the aces dealt) will be in hands that contain two aces. But the conditional probability represented by these numbers is the *probability that an ace will be dealt into a hand containing two aces, given that it is dealt at all.* This is not the same as the *probability that a hand contains two aces, given that it contains one.* The problem posed by Copi (1968) has to do with the latter probability and not the former. The proof of the pudding here is in the eating. The individual who believes that the probability that a hand contains two aces, given that it contains one, is 3/7 and accepts betting odds consistent with this belief will lose money to one who bets in accordance with the belief that that conditional probability is 3/11.

A. I. Dale (1974), like Rose, fails to note the ambiguity in Copi's statement of the problem, and tacitly assumes that the friend announces that the hand contains an ace whenever it contains at least one ace. This is seen in the assertion that "Copi's question may equivalently be phrased as 'find the probability that both cards are aces, *given that at least one of them is an ace*'" (p. 205). But one of the main points I am trying to make in this discussion is that such a rephrasing of the problem requires the making of an assumption, and that the overlooking of this fact is at the base of much of the misunderstanding about conditional probabilities. But again, suppose we explicily make the assumption that Rose and Dale both implicitly make.

A. I. Dale (1974) criticizes Rose's (1972) analysis on the grounds, in part, that it considered the order of the cards in a hand to be irrelevant to the solution of the problem: "The order is of supreme importance, and failure to recognize this fact has, I think, been responsible for wrong solutions to both the problem under consideration here and many others" (p. 204). Dale's analysis begins with a listing of all 56 possible two-card hands, order being taken into account, and proceeds with a consideration, by enumeration, of specific conditional probabilities. The probability that the second card is an ace, given that the first one is an ace, for example, is found by counting the number of instances of the first card being an ace (28) and the number of *those* cases for which the second card is also an ace (12) and taking the ratio, which, in this case, is 3/7. Similarly, to find the probability that both cards are aces, given that at least one is, one counts the number of hands that contain at least one ace (44) and the number of those that contain two aces (12) and take the ratio, which is 3/11. (Although this ratio was obtained by considering the full set of 56 hands, taking card order into account, the same 3/11 would be obtained by enumaration using only the 28 possibilities when card order is not taken into account.)

A. I. Dale (1974) attributes Rose's answer of 3/7 to the question of the probability that both cards are aces, given that one is, to "the failure to notice that the event 'at least one card is an ace' is the union of the two mutually-exclusive events 'exactly one card is an ace' and 'exactly two cards are aces'" (p. 205). Dale goes on to argue as follows: "I do not think it is correct to regard any one of the aces in the above array [Dale's listing of all 56 possible hands] as providing the same amount of support for the information that there is at least one ace. That is, while the hand A_s, K_h, (say), provides a certain amount of support for this proposition, the hand A_s, A_h does not provide *more* support, which is what Rose seems to be claiming" (p. 205). I confess to not understand exactly what is being said here, so I do not see it as a convincing account of Rose's miscalculation, though miscalculation I consider it to be.

Both Rose and Dale accept Copi's answer of 3/7 as the probability that the hand contains two aces when the friend announces that he has the ace of spades. We should note that this too involves an unstated assumption, namely that the friend announces that his hand contains an ace of spades whenever it does so. If, in those instances in which he held an ace of spades, he would sometimes announce holding the *other* card, the probability of holding two aces, given the announcement of holding an ace of spades, would not necessarily be 3/7. What it would be would depend on the specifics of the conditions under which the ace of spades would be announced.

In a commentary on the analyses by Rose and Dale, Faber (1976) puts his finger on the crux of the matter on which this discussion has focused, namely the indeterminacy of Copi's problem in the absence of knowledge, or an assumption about, the conditions under which the friend (the player in Faber's terms) announces that his hand contains an ace (or that it contains a specified ace). "To calculate probabilities it is not enough to know," Faber argues, "that there is at least one ace in the hand. We must also know how this evidence came to light; that is we must know the rule by which the player decided what to announce" (p. 284). Faber considers three possibilities:

"Rule 1: The player chooses randomly which of the two cards to inform us about."

"Rule 2: The player is obliged to announce an ace when he holds one, but is free to specify the suit of either of two aces when two are present."

"Rule 3: The player must announce an ace if he holds one, and must specify the spades ace in preference to another suit when two aces are present" (pp. 284, 285).

Faber (1976) argues that the probability that the hand contains two aces given that the player announces that it contains the ace of spades, or just that it contains an ace, under each of these rules is as shown in Table 5.6.

Faber (1976) suggests that, in the absence of any hints from Copi (1968) about the rule under which the friend in the original statement of the problem was operating, it is reasonable to assume Rule 1, and he argues that under this assumption, the two probabilities that Copi claims are different are really the same.

As I understand Faber's comments, he considers the situation in which the player invariably announces holding a particular card by identifying *both* its face and suit—ace of spades, king of hearts; he does not consider the situation in which the player is free sometimes to announce holding an ace (or a king) *without* naming the suit. The rules, as expressed earlier, do not specify this constraint, but Faber's analysis suggests that this was the intention. Moreover, the constraint is essential to preclude perverse but informative implementations of the rules, as expressed. For example, Rule 3, would permit the player to announce having an ace, without reporting its suit, whenever he holds only one, and to specify the ace of spades *only* when he holds *two* aces, one of which is the ace of spades; this reporting rule would, of course, provide useful information to an observer who knew it was being used. In general, any implementation of the rules, as expressed, in which the player decides, on the basis of a card's face, whether or not to report its suit could be informative in a similar way. The following comments are predicated on the assumption that the rules, as intended by Faber, could be expressed as follows.

Rule 1: The player chooses randomly which of the two cards to inform us about, and reports the face and suit of that card.

TABLE 5.6

Probability That the Hand Contains Two Aces, Contingent on What the Player Announces, Under Each of Three Announcement Rules Accoring to Faber (1976)

	Player Announces	
	The Ace of Spades	*An Ace Other Than Spades*
Rule 1	3/7	3/7
Rule 2	3/11	3/11
Rule 3	3/7	3/11

Rule 2: The player is obliged to announce an ace and its suit when he holds one, and is to select which suit to report at random when he holds two aces.

Rule 3: The player must announce an ace, reporting its suit, if he holds only one, and must specify the spades ace in preference to another suit when he holds two aces, one of which is the ace of spades.

Recall that of the 28 possible hands (ignoring order), 6 have two aces, 6 have two kings, and 16 have an ace and a king. It may help at this point to lay out the hands to highlight these facts. This is done in Table 5.7. The hands that contain the ace of spades are in bold print. We assume, of course, that all possible hands are equally probable.

Under Rule 1, when the player gets a particular one of these hands, he is equally likely to reveal the identity of either card. So when he gets an ace and a king, he is equally likely to announce the one as the other, and when he gets two aces, he is equally likely to name either suit. The probability of getting two aces in hand is 6/28. In only three of such cases will one of the two aces be the ace of spades, and on only 1/2 of those instances will the player announce "ace of spades"; so the joint probability that the player's hand will contain two aces *and* that he will announce "ace of spades," *p*(AA&"A_s"), is (3/28)(1/2) = 3/56. The probability that the hand will contain the ace of spades is 7/28 and, inasmuch as the player will announce "ace of spades" on half of such instances, the probability that the player will announce "ace of spades," *p*("A_s"), is (7/28)(1/2) = 7/56. So the probability that the player has two aces, given that he announces having the ace of spades, *p*(AA | "A_s"), is

TABLE 5.7

All Possible Two-Card Hands (Ignoring Order) That Can Be Dealt From a Deck Consisting of Four Aces and Four Kings

A_sA_h	**A_sK_s**	A_dK_s	K_sK_h
A_sA_d	**A_sK_h**	A_dK_h	K_sK_d
A_sA_c	**A_sK_d**	A_dK_d	K_sK_c
A_hA_d	**A_sK_c**	A_dK_c	K_hK_d
A_hA_c	A_hK_s	A_cK_s	K_hK_c
A_dA_c	A_hK_h	A_cK_h	K_dK_c
	A_hK_d	A_cK_d	
	A_hK_c	A_cK_c	

$$p(\text{AA} \mid \text{“A}_s\text{”}) = p(\text{AA\&“A}_s\text{”})/p(\text{“A}_s\text{”}) = (3/56)/(7/56) = 3/7.$$

Now consider the probability that the player has two aces, given that he announces having an ace other than the ace of spades (still under Rule 1). Again the probability of getting two aces in hand is 6/28. In three such cases one of the two aces will be something other than spades, and in the other three both of them will be. In half of the former cases and all of the latter, the player will announce an ace other than spades, so the joint probability that the player's hand will contain two aces *and* that he will announce a nonspades ace, p(AA&“$\text{A}_{\text{not-s}}$”), is (3/28)(1/2) + (3/28) = 9/56. The probability that the hand will contain an ace other than spades is 18/28; in three of such hands both cards are nonspades aces and in the remaining 15 only one is, so he will announce a nonspades ace in three cases and in half of the other 15, which is to say the probability that the player will announce a nonspades ace, *p*(“$\text{A}_{\text{not-s}}$”) is (3/28) + (15/28)(1/2) = 21/56. So the probability that the player has two aces, given that he announces having a nonspades ace, *p*(AA | “$\text{A}_{\text{not-s}}$”), is

$$p(\text{AA} \mid \text{“A}_{\text{not-s}}\text{”}) = p(\text{AA\&“A}_{\text{not-s}}\text{”})/p(\text{“A}_{\text{not-s}}\text{”}) = (9/56)/(21/56) = 3/7.$$

A similar analysis of the implications of Rule 2 will show that the corresponding probabilities are both 3/11. Consider first the case in which the player announces “ace of spades.” Given that the player is obliged to announce an ace when he has one, but is free (let us assume equally likely) to specify the suit of either ace when two are present, the joint probability that the player's hand will contain two aces *and* that he will announce “ace of spades,” is, as with Rule 1 (3/28)(1/2) = 3/56. Again, the probability that the hand will contain the ace of spades is 7/28, but now the player will announce “ace of spades” whenever the ace of spades is paired with a king and on half of the instances when it is paired with another ace, so the probability that the player will announce “ace of spades” is (4/28) + (3/28)(1/2) = 11/56. Thus the probability that the player has two aces, given that he announces having the ace of spades, is

$$p(\text{AA} \mid \text{“A}_s\text{”}) = p(\text{AA\&“A}_s\text{”})/p(\text{“A}_s\text{”}) = (3/56)/(11/56) = 3/11.$$

Suppose, again under Rule 2, that the player announces having an ace other than the ace of spades. The joint probability that the player's hand will contain two aces *and* that he will announce an ace other than spades is (3/28) + (3/28)(1/2) = 9/56. The probability that the hand will contain an ace other than spades is 18/28; the ace of spades will be one of the aces in three of these cases but not on the other 15, so the player will announce an ace other than spades with probability (15/28) + (3/28)(1/2) = 33/56. So in this case, the probability that the player has two aces, given that he announces having a nonspades ace, is

$$p(\text{AA} \mid \text{“A}_{\text{not-s}}\text{”}) = p(\text{AA\&“A}_{\text{not-s}}\text{”})/p(\text{“A}_{\text{not-s}}\text{”}) = (9/56)/(33/56) = 3/11.$$

Finally, let us consider Rule 3, which obliges the player to announce that his hand contains the ace of spades whenever it does so. With this rule, the joint probability that the player's hand will contain two aces *and* that he will announce "ace of spades," is 3/28. The probability that the player will announce "ace of spades" is the same as the probability that the hand contains the ace of spades, or 7/28. So the probability that the player has two aces, given that he announces having the ace of spades, is

$$p(\text{AA} \mid \text{“A}_{\text{s}}\text{”}) = p(\text{AA\&“A}_{\text{s}}\text{”})/p(\text{“A}_{\text{s}}\text{”}) = (3/28)/(7/28) = 3/7.$$

Rule 3 is a bit more complicated when applied to those cases in which the player announces some ace other than spades. This is because a listener who knows the rule knows that when the player announces a nonspades ace, the other card in the hand is *not* the ace of spades; if it were the ace of spades he could not have announced the nonspades ace. The joint probability that the player's hand will contain two aces *and* that he will announce an ace other than spades is 3/28. Inasmuch as there are 15 hands that contain one or two aces, neither of which is the ace of spades, the probability that he will announce an ace other than spades is 15/28. So given rule 3 as expressed on page 166, the probability that the player has two aces, given that he announces having a nonspades ace, is

$$p(\text{AA} \mid \text{“A}_{\text{not-s}}\text{”}) = p(\text{AA\&“A}_{\text{not-s}}\text{”})/p(\text{“A}_{\text{not-s}}\text{”}) = (3/28)/(15/28) = 1/5.$$

There are at least two other rules by which the player of the game described by Copi (1968) could decide what to announce, both of which are plausible possibilities of what Copi had in mind and are unambiguous:

Rule 4: The player must announce that he holds at least one ace if he does so, *without* specifying its suit if he holds only one and without specifying the suit of either if he holds two, and he must announce nothing else.

Rule 5: The player must announce the ace of spades if he holds it, and he must announce nothing else.

With Rule 4, an observer always learns whether or not a hand contains at least one ace; with Rule 5, one always learns whether or not it contains the ace of spades. With Rule 4, the joint probability that the player's hand will contain two aces *and* that he will announce "at least one ace" is 6/28. The probability that the player will announce "at least one ace" is the same as the probability that the hand contains at least one ace, or 22/28. So the probability that the player has two aces, given that he announces having at least one ace, is

$$\begin{gathered} p(\text{AA} \mid \text{"at least one ace"}) \\ = p(\text{AA\&"at least one ace"})/p(\text{"at least one ace"}) \\ = (6/28)/(22/28) = 3/11. \end{gathered}$$

With Rule 5, the joint probability that the player's hand will contain two aces *and* that he will announce "the ace of spades" is 3/28. The probability that he will announce "the ace of spades" is the same as the probability that the hand contains the ace of spades, or 7/28. So the probability that the player has two aces, given that he announces having the ace of spades, is

$$p(\text{AA} \mid \text{"A}_s\text{"}) = p(\text{AA\&"A}_s\text{"})/p(\text{"A}_s\text{"}) = (3/28)/7/28) = 3/7.$$

Perhaps these two probabilities represent the distinction that Copi (1968) had in mind in claiming that the probability that both cards are aces is different when the holder of the hand announces that one of the cards is an ace than when he announces that one of them is the ace of spades; what was missing from Copi's claim was a recognition that in order to determine the probabilities, it is not enough to assume the card holder's report is truthful, but an assumption must be made about the conditions under which he will make various possible (truthful) reports. Faber's (1976) critique of Copi's claim was justified in pointing this out, but the rules he articulated were themselves somewhat ambiguous, and, given unambiguous restatements of them, his anlysis to illustrate the effects of different possible reporting rules was not quite right with respect to Rule 3.

In the article following that of Faber in the same journal, S. Goldberg (1976) comments on the problem posed by Copi and on the earlier commentaries by Rose, Dale, and Faber. Goldberg chooses to view Copi's problem as "most reasonably formulated as simply asking us to compute the conditional probabilities that both cards in the hand are aces, given that the hand contains at least one ace, and then given that the hand contains the ace of spades" (p. 287). How we come to know that the hand contains at least one ace, or that it contains the ace of spades, "whether by announcement or otherwise, is irrelevant," Goldberg argues, and raising this issue unnecessarily complicates the picture. It should be clear from the forgoing that, given this interpretation of the problem, the answer is 3/11 in the former case and 3/7 in the latter. But should we accept this interpretation? Failure to recognize that inferring the probability of an event conditional on the report of some observation may require an assumption about the conditions under which the report would be made—and that different assumptions could justify different inferences—is precisely the point at issue. And the literature provides many evidences that lack of clarity on this point can lead to considerable confusion and debate.

S. Goldberg (1976) argues that "the two announcements in Copi's exercise are most reasonably and simply interpreted as truthful affirmative responses to the questions: 'Does the hand contain at least one ace?' and 'Does the hand contain the ace of spades?'" (p. 287). (Rules 4 and 5 are designed to yield answers to these questions.) This is an opinion, and perhaps one that many people share, or would endorse if asked. But not everyone who has thought about the problem shares this opinion, as the foregoing discussion illustrates. Moreover, if what one wants to get at is people's understanding of the theory of probability by seeing how well they can solve problems, the problems should be posed in terms that are open to as little opinion-based interpretation as possible. Copi's problem could easily be stated as:

> Imagine the following situation. You have a deck of cards from which all but the four aces and four kings have been removed. From this abbreviated deck, you deal two cards, at random, to a friend. You then ask if the dealt hand contains at least one ace, and your friend answers truthfully that it does; what is the probability that both her cards are aces? Now suppose you deal two cards, at random, from the same eight-card deck, and this time you ask your friend whether the hand contains the ace of spades. Again, assume she answers truthfully that it does; what is the probability then that both her cards are aces?

This statement of the problem is not ambiguous, in my view, and an analysis will readily show that the answer to the first question is 3/11, whereas the answer to the second one is 3/7. But the literature is full of probability problems that are posed in ambiguous terms. One can only assume that the ambiguity generally has not been recognized as such by the problem posers, and the many conflicting accounts of what the solutions are attest to the ease with which different interpretations are made.

TWO ACES AND A JACK

As a further illustration of how easy it is to state probability problems that are indeterminate in the absence of assumptions beyond what is given in the problem statement, I offer the following from Gillman (1992):

> [Consider] a deck of four cards, two aces and two jacks, from which you are dealt a hand of two cards. There are six possible hands, one of them consisting of the two aces, so the probability you have both aces is 1/6. If it is given that the hand contains an ace we have eliminated the two jacks, and the probability for both aces goes up to 1/5. But if it is given that you have the ace of hearts, then your other card is either the ace of spades or one of the jacks, and the probability that you are holding both aces is now 1/3.
>
> Carrying this to the extreme, consider a two-card hand from a deck of *three* cards, two aces and a jack. There are three possible hands, and the probability that you have the two aces is 1/3. If you state that the hand contains an ace, I smirk. But if we are given that the hand contains the ace of hearts, the probability for both aces goes up to 1/2. At this point (if not long since) your friend enters the picture with a "proof" that the probability of both aces is 1/2, with or without any condition: "You have an ace. Either it is the ace of hearts or the ace of spades. If it is the ace of hearts, then as we have just proved, the probability of both aces is 1/2. If it is the ace of spades, then, similarly, the probability for both aces is 1/2. So in either case, it is 1/2. So it is 1/2." It is easier to detect the flaw in this reasoning than to get your friend to understand it. (p. 6)

I think that, in the absence of an assumption, the situation is not as clear as Gillman (1992) seems to suggest. The assumption that is required has to do with the conditions under which we are given the knowledge that the hand contains the ace of hearts. Did someone look at the hand and report that it contained the ace of hearts? Were the rules of reporting such that we were sure to have been informed of the presence of the ace of hearts if one were there? Or did they permit the reporting of the ace of spades, instead of the ace of hearts, if the hand contained both aces?

Focusing first on the three-card case, there are at least three plausible rules to consider: (a) Select at random which card to report, and report its face and suit, (b) report only whether or not the ace of hearts is in the hand, and (c) report the suit of the ace when there is only one ace and the suit of a randomly selected ace when there are two of them. The three situations are shown in Tables 5.8 through 5.10. The probability that the second card is an ace, given that we have been informed that one of them is the ace of hearts is 1/2 with each of the first two rules but 1/3 with the third. So again, the conditional probability that the hand contains two aces, given the knowledge that it holds the ace of hearts depends on what we assume about the rule by which that knowledge was obtained. This is equally true in the four-card case. I will not work through alternative possible assumptions, but perhaps it is apparent that the answer one gets will differ depending on whether the assumption one makes assures that the ace of hearts will be reported for all hands that contain it or permits the possibility that it will not, say by permitting the reporting of the ace of spades

TABLE 5.8

The Rule Is to Select at Random Which Card to Report

	Card Reported			
Hand	A_H	A_S	J	Σ
A_HA_S	1/6	1/6	0	1/3
A_HJ	1/6	0	1/6	1/3
A_SJ	0	1/6	1/6	1/3
Σ	1/3	1/3	1/3	1

TABLE 5.9

The Rule Is to Report Only Whether or Not the Hand Contains the Ace of Hearts

	Report		
Hand	*Yes*	*No*	Σ
A_HA_S	1/3	0	1/3
A_HJ	1/3	0	1/3
A_SJ	0	1/3	1/3
Σ	2/3	1/3	1

TABLE 5.10

The Rule Is Always to Report the Suit of the Ace When There Is Only One Ace and the Suit of a Randomly Selected Ace When There Are Two

	Card Reported		
Hand	A_H	A_S	Σ
A_HA_S	1/6	1/6	1/3
A_HJ	1/3	0	1/3
A_SJ	0	1/3	1/3
Σ	1/2	1/2	1

(rather than the ace of hearts) on some proportion of the cases in which the hand contains both aces.

A similar argument applies to the four-card case in which it is "given" that the hand contains an ace without anything being said about suit. Again what we make of this depends on what we assume about the conditions for obtaining this information. If we assume that we are told there is an ace whenever there is at least one ace, the probability that there are two aces conditional on being informed that there is one, is, as Gillman (1992) claims, 1/5. But if we assume that when there is one ace and one jack, we are equally likely to be told that there is one jack as we are to be told that there is one ace, the probability that there are two aces, conditional on being informed that there is one, is 1/3. This follows from the fact that we will be informed that there is an ace in all instances in which the hand contains two, but in only half those of the four times as many instances in which it contains only one.

Gillman was not insensitive to the issue that I have been trying to address. The card problem was adapted from one described by Ball (1892) (now Ball & Coxeter, 1987). Gillman (1992) used the term "given that" a hand contains specified cards, rather than saying that someone has *asserted* that it does so, as Ball had done, for the explicit purpose of avoiding the ambiguity I have been claiming. To interpret an assertion, Gillman notes, it would be necessary to know how one decided what to assert: "My present rule is that you are to state whether your hand contains an ace" (p. 6). With that interpretation of "given," I agree with Gillman's figures. Gillman also considers the possible rule, in the case of the four-card problem, of reporting one of the cards at random and gives 1/3 as the probability of two aces conditional on knowledge of the presence of one in this case.

Essentially the same problem (using aces and deuces of spades and clubs) was discussed by Freund (1965), who argued that the problem as he found it—in a book of mathematical puzzles by Gamow and Stern (1958)—was incomplete and could not be solved, at least not "without smuggling in unwarranted assumptions" (p. 44). Freund identified as critical the question of how we come to know that the hand in question (which in his scenario is held by an opponent) contains at least one ace, or a specified ace. To put the problem in perspective, he proposes imagining that a spy provides us with the required knowledge, and considers two possible ways in which the spy gets information:

> Case 1: The spy looks at our opponent's entire hand; either he reports whether or not he sees (at least) one ace, or he also reports the suit (flipping a coin to decide whether to report spades or clubs when he sees both aces in our opponent's hand).

Case 2: The spy has the chance to see only one card (randomly selected from our opponent's hand) and he either reports whether or not it is an ace, or he also reports the suit. (p. 29)

Here is Freund's (1965) analysis of Case 1:

> If he merely reports that he sees (at least) one ace, we find ourselves in the first situation described above [elimination of the one hand that has no aces, leaving five equiprobable hands]. Using Bayes' Rule it can easily be shown that the 5 remaining hands have equal *a posteriori* probabilities and, hence, that the correct odds are 4 to 1. If he also reports the suit, namely, that he sees the ace of spades, the use of Bayes' Rule shows that the 3 remaining hands do *not* have equal *a posteriori* probabilities. As a matter of fact, the *a posteriori* probability of the hand which consists of both aces is only half of that of the other two remaining hands, and it follows that the correct odds against our opponent having both aces are still 4 to 1. The same argument applies also if the spy reports the ace of clubs. (p. 29)

I think this analysis is correct, given an unstated assumption that I will make explicit. The wording "he reports *whether or not* he sees (at least) one ace, or he *also* reports the suit ..." [emphasis added] makes it clear, in my view, that the spy's rule is to report seeing an ace *in all cases in which he does see at least one.* This precludes the possibility that, upon seeing an ace and a deuce, he will report seeing the deuce and *not* report seeing the ace. And the proviso that, if reporting suits, the spy will flip a coin to decide which ace to report when he sees two of them rules out the possibility of a bias in reporting suits when he has a choice of which ace to report. So far, so good; but there remains an ambiguity that needs to be resolved, and thus the need for a further assumption.

The assumption is needed because Case 1 can be interpreted in two ways. One way to interpret it is as a description of two subcases: Case Ia, in which the spy *always* reports *only* whether or not he sees (at least) one ace and does *not* report the suit, and Case Ib, in which he *always* reports the suit of an ace if he sees at least one, flipping a coin to decide which one to report when he sees two (flipping the coin in private, of course, so as not to reveal by this act the fact that the hand contains two aces). Given this interpretation of Case I, Freund's (1965) analysis, I believe, is correct. However, another interpretation of Case I does not rule out the possibility that the spy *sometimes* reports *only* whether or not he sees an ace without reporting the suit, and *sometimes* reports *both* the presence of an ace and its suit (or the suit of a randomly chosen one of them if he sees two).

This interpretation does not preclude the operation of an information-conveying bias such as would pertain, for example, if the spy reported seeing an ace whenever the hand contained at least one ace and, for those hands that contained only one ace, reported the ace's suit only, or primarily, when it was the ace of spades. This seems an unlikely interpretation, but it is a possible one, and if the spy were motivated to give a deceptive, albeit accurate, report, this is one way he could do it. In dealing with spies, it is well to be aware of all the possibilities! So the correctness of Freund's (1965) analysis of Case 1 rests, I claim, on an unstated assumption. The assumption could be that Case 1 is to be interpreted as two subcases, as described earlier, or it could be that the spy only sometimes reports the suit of an ace in a one-ace hand, but his decision of when to do so is independent of which ace the hand contains.

Given either of these assumptions, Freund (1965) is correct in giving 4 to 1 as the odds against the hand having two aces, given the spy's report that it has at least one, or his report that it has the ace of spades. This may be easier to see in the first instance than in the second. In the first instance, the spy is equally likely to report seeing at least one ace in five of the six possible hands, only one of which contains two aces, so the probability that the hand contains two aces, given the report that it contains at least one, is 1/5. The instance in which the spy reports the suit is slightly more complicated. Three of the six possible hands contain the ace of spades, and one of these three contains two aces. The important thing to notice is that, given that the spy is to flip a coin to decide which ace to report when the hand contains two aces, he is only half as likely to report the ace of spades for the two-ace hand as he is to report the ace of spades for either of the hands that contain the ace of spades as the only ace, so the probability that the hand contains two aces, given that the spy reports that it contains the ace of spades is $(1/2)/[(1/2) + 1 + 1] = 1/5$, or, if you prefer, $(1/2)(1/6)/[(1/2)(1/6) + (1/6) + (1/6)] = 1/5$. The 1/6 of the latter equation represents the a priori probability of each hand containing the ace of spades, and the 1/2 represents the fact that that ace would be reported for only half of the hands that contain both aces.

Freund's (1965) analysis of Case 2 yields odds of 2 to 1 against the hand having both aces, given the spy's report that the card he sees is an ace, and independently of whether he reports the suit of that ace. Case 2, as described by Freund, is ambiguous in the same way as Case 1. One interpretation of it is as two subcases: In Case 2a, the spy *always* reports *only* whether or not the card he inspects is an ace; in Case 2b, he *always* reports whether or not it is an ace *and* gives the suit if it is. (He may report the suit if it is not an ace as well; this makes no difference.) An alternative interpretation leaves open the possibility that the spy sometimes reports only whether or not the card he sees is an ace without reporting the suit, and sometimes reports both whether it is an ace and its suit if it is. As in Case 1, the second interpretation does not rule out the possibility of the spy making the probability of reporting the suit of an ace contingent on which ace he sees. In Case 2, however, such a bias would not convey useful information; if the spy only *sees* one of the cards, his honest report of whether or not it is an ace is informative with respect to the a posteriori conditional probability that the other card is also an ace, but his report (or lack of report) of that ace's suit is not.

The 2-to-1 odds against the hand having both aces, given the spy's report that the card he sees is an ace, or that it is the ace of spades, is easily seen if we consider all 12 two-card hands that are possible when card order is taken into account. Seeing only one card is equivalent to looking always at the first card dealt (left) or always at the second card dealt (right). As shown in Table 5.11, two out of six of the hands that have an ace in the left (right) position have one also in the right (left) position. And one out of three of the hands that have an ace of spades in the left (right) position have an ace in both positions, which is what Freund's (1965) analysis claims.

SISTERS AND BROTHERS

The following problem was given to me in correspondence by Ruma Falk: "In a large random sample of men and women, should we expect the men to have more sisters than women? And for men to have more sisters than brothers?" Falk's answer to both questions was no.

I am one of four children—two boys and two girls. My brother and I each have one brother and two sisters; each of my sisters has one sister and two brothers. It seemed obvious to me that, on average, men would have more sisters than do women, as well as more sisters than brothers. My reasoning was that this must be the situation in all families with an equal number of boys and girls and that any departure from this rule in one direction among families with more boys than girls would be balanced by an opposite departure among families with more girls than boys.

Before writing Falk to tell her I thought her answer to the question—that men are expected to have the same number of sisters as do women and the same number of sisters as brothers—to be wrong, I decided to work things out for

TABLE 5.11
All Possible Two-Card Hands Dealt From Four Cards (Aces and Deuces of Spades and Clubs) Taking Order Into Account

A_sA_c	A_cD_s	A_cA_s	D_sA_c
A_sD_s	A_cD_c	D_sA_s	D_cA_c
A_sD_c	D_sD_c	D_cA_s	D_cD_s

several family sizes. The results of this exercise are shown in Tables 5.12, 5.13, and 5.14 for families with two, three, and four children respectively.

Each row of each table represents one of the possible combinations of males and females in a family of the specified number of children. Column C indi-

TABLE 5.12

Representing a Family of Two Children

A	*B*	*C*	*D*	*E*	*F*	*G*				
Males	*Females*	*Families*	*M's Bs*	*M's Ss*	*F's Bs*	*F's Ss*	*ACD*	*ACE*	*BCF*	*BCG*
2	0	1	1	0	0	0	2	0	0	0
1	1	2	0	1	1	0	0	2	2	0
0	2	1	0	0	0	1	0	0	0	2

TABLE 5.13

Representing a Family of Three Children

A	*B*	*C*	*D*	*E*	*F*	*G*				
Males	*Females*	*Families*	*M's Bs*	*M's Ss*	*F's Bs*	*F's Ss*	*ACD*	*ACE*	*BCF*	*BCG*
3	0	1	2	0	0	0	6	0	0	0
2	1	3	1	1	2	0	6	6	6	0
1	2	3	0	2	1	1	0	6	6	6
0	3	1	0	0	0	2	0	0	0	6

TABLE 5.14

Representing a Family of Four Children

A	*B*	*C*	*D*	*E*	*F*	*G*				
Males	*Females*	*Families*	*M's Bs*	*M's Ss*	*F's Bs*	*F's Ss*	*ACD*	*ACE*	*BCF*	*BCG*
4	0	1	3	0	0	0	12	0	0	0
3	1	4	2	1	3	0	24	12	12	0
2	2	6	1	2	2	1	12	24	24	12
1	3	4	0	3	1	2	0	12	12	24
0	4	1	0	0	0	3	0	0	0	12

cates the number of families, on average (from totals of 4, 8, and 16 in Tables 5.12 through 5.14, respectively), that would be expected to have that combination. This column contains, of course, the binomial coefficients. Columns D and E give the number of brothers and sisters a male would be expected to have; F and G give the same information for females. The four rightmost columns give, respectively, for each combination of males and females in the family, the numbers of brothers of males (ACD), sisters of males (ACE), brothers of females (BCF), and sisters of females (BCG).

Adding over rows (male–female combinations) shows the total expected numbers of brothers and sisters to be the same for males and females. Considering all males in families of three children, 3/12, or 1/4, are in families of three males (and so have two brothers), 6/12, or 1/2, are in families of two males (one brother), and 3/12, or 1/4, are in families of one male (no brothers). So the average number of brothers for males in families of three children is (1/4)(2) = (1/2)(1) + (1/4)(0) = 1. Similar calculations show that the average number of sisters for males (as well as brothers for females and sisters for females) in families of three children is also (1/4)(2) + (1/2)(1) + (1/4)(0) = 1. The same analysis will show equivelance for families of other sizes as well.

Another way to approach the problem is to ask: Given that one is a boy (girl) in a family of n children, what is the expected number of brothers and the expected number of sisters he (she) will have? The possibilities for $n = 2$ are as shown in Table 5.15, taking birth order into account.

The total number of boys represented in the table is four and the total number of brothers of boys [Σ boys × (bros/boy)] is two, so the expected number of brothers per boy is .5. Similarly, the total number of sisters of boys is two, so the expected number of sisters per boy in a family of two is also .5. These numbers are consistent with the expected number of siblings, which, of course, must be one. The same figures hold for a girl in a family of two children.

TABLE 5.15

Possible Combinations of Brothers and Sisters in a Family of Two Children, Taking Birth Order Into Account

	boys	*bros/boy*	*sis/boy*	*girls*	*sis/girl*	*bros/girl*
BB	2	1	0	0	0	0
BG	1	0	1	1	0	1
GB	1	0	1	1	0	1
GG	0	0	0	2	1	0

Expansion of Table 5.15 to represent families of three and four children shows—no surprise now—that in a family of three, the expected number of brothers per boy, and per girl, and the expected number of sisters per boy and per girl are 1.0 in all cases, and that in a family of four, the comparable expected number is 1.5 in all cases. Table 5.16 gives the details for a family of three children.

Having worked out these examples, I became convinced of the truth of Falk's claim that males would have as many brothers as sisters and as many brothers as would females, fortunately before sending off my letter to tell her I believed her conclusion to be wrong. So instead I sent her this analysis; she, with great diplomacy, pointed out to me that asking a boy or girl selected at random from families of *n* children how many brothers or sisters he or she has is essentially equivalent to asking a person selected at random from families of *n* – 1 children how many boys or girls it has, and that if one sees this equivalence, such an analysis is not necessary. The reasoning is spelled out in Falk (1993, pp.162–164) and Falk and Konold (1992).

SUMMARY

One might expect that when the theory was just beginning to be developed, being surprised by results was not an uncommon experience, and that disagreement among mathematicians as to how probabilities should be computed was not rare. Several accounts of specific surprises and disagreements are recorded in the literature. One such account relates to the surprise of Chevalier de Méré upon learning that although the probability of obtaining at least one six in four tosses

TABLE 5.16
Possible Combinations of Brothers and Sisters in a Family of Three Children, Taking Birth Order Into Account

	boys	*bros/boy*	*sis/boy*	*girls*	*sis/girl*	*bros/girl*
BBB	3	2	0	0	0	0
BBG	2	1	1	1	0	2
BGB	2	1	1	1	0	2
GBB	2	1	1	1	0	2
BGG	1	0	2	2	1	1
GBG	1	0	2	2	1	1
GGB	1	0	2	2	1	1
GGG	0	0	0	3	2	0

of a die is greater than 1/2, the probability of obtaining at least one double-six in 24 tosses of a pair of dice is not (Freudenthal, 1970). Apparently de Méré had reasoned that inasmuch as there are 36 possible outcomes of the toss of two dice and 24:36 = 4:6, the probability of obtaining at least one double-six in 24 tosses of two dice should be the same as getting at least one six in four tosses of a single die (Falk, 1992). In fact the latter probability is .518 and the former .491.

The great 18th-century mathematician, D'Alembert, believed the probability of producing at least one head in two tosses of a coin to be 2/3 rather than 3/4, on the grounds that the game is finished with the first toss when it produces a head, which implies a sample space of three outcomes: H, TH, and TT; and he maintained this position despite arguments of his contemporaries to the contrary (Todhunter, 1865/2001). D'Alembert's mistake, according to contemporary views of probability, was in assuming these three outcomes to be equiprobable when, in fact, the first is twice as likely as each of the others. Other examples of differences in intuitions about probability among the early developers or users of probabilistic concepts could be given.

Probability theory is now a well-established branch of mathematics, but students, and perhaps even experts, still find some results surprising and counterintuitive, at least when first considered. The problems considered here illustrate some of the subtleties that can be involved in probabilistic reasoning. Other problems illustrating similar subtleties have been discussed by Mosteller (1965) and Falk (1993). Difficulties in probabilistic reasoning are also illustrated by situations that are sometimes referred to as paradoxes and dilemmas, the subjects of the following chapter.

CHAPTER 6

Some Probability Paradoxes and Dilemmas

PARADOXES

What constitutes a paradox is, to some degree, a matter of semantics. One dictionary definition is "a statement that is seemingly contradictory or opposed to common sense and yet is perhaps true" *(Webster's New Collegiate Dictionary);* there are others. Paradoxes that proved to be so important to the history of mathematics, especially during the early part of the 20th century, often involved self-contradictory statements, or what appeared to be different but equally valid mathematical proofs leading to contradictory conclusions.

Many problems involving probability theory have been described and discussed in the literature as paradoxes. Some of them can be resolved readily by carefully analyzing the situations involved; others are less easily dispatched. Sometimes the same problem is referred to by some authors as a paradox and by others as something else. The problem of the inquisitive prisoner is a case in point (for a description, see Nickerson, 1996); it has been called a paradox

(Weintraub, 1988; Zabell, 1988), a dilemma (Mosteller, 1965), and an absurdity (Székely, 1986).

I will not attempt a precise definition here. The following problems have been called paradoxes, or are similar to others that have been so called. Readers who feel that some of them are better described by other words should make what they consider to be suitable substitutions in terminology.

The St. Petersburg Paradox

Possibly the most famous paradox involving probability theory is one described by Nicholas Bernoulli in a letter to Pierre de Montmort who later published it in a book on games of chance (Montmort, 1708). Its name comes from the fact that Nicholas' cousin, Daniel Bernoulli, published his resolution of it (not universally accepted as a resolution) in the annals of the Academy of St. Petersburg in 1738.

The "expected value" of the outcome of a probabilistic event is said to be the sum of the products of the values of each of the possible outcomes multiplied by its probability of occurrence. In other words

$$E = \sum_{i=1}^{n} p_i V_i \, ,$$

where p_i and V_i are the probability and value, respectively, of the *i*th outcome. Thus, if you stood to gain \$1 by the toss of a head with a fair coin and \$0 with the toss of a tail, the expected value of a toss would be (.5 × \$1) + (.5 × \$0) or 50 cents. Presumably, if you were offered the chance to purchase this gamble, you should consider any purchase price less than \$.50 to be a good buy and any price over \$.50 to be a poor one.

Consider the following gamble. A fair coin is to be tossed until it comes up head; that is to say the tossing will continue as long as the outcomes are tails, the first head terminates the game, and when it occurs determines how much you win. (As initially posed by N. Bernoulli to Montmort, the problem involved the tossing of dice, but shortly later Cramer restated the problem as a game of heads and tails [Jorland, 1987], and the latter version has generally been discussed since then.) If the first toss is head, the game is over and you receive \$2. If a tail occurs first and then a head, you receive \$4; if a tail occurs twice before the first head, you receive \$8; if three times, \$16, and so on, the amount you win doubling with each additional toss of a tail; so if a head occurs first on the *k*th toss, you win 2^k dollars. In general the situation is as follows:

Possibility	*Probability*	*$ won*
H	1/2	2
TH	1/4	4
TTH	1/8	8
TTTH	1/16	16
•	•	•
•	•	•
T ... (k – 1 times) H	$1/2^k$	2^k

What would you consider a reasonable price to pay for the opportunity to play this game? The expected value of the gamble, being the sum of the products of all possible winnings and their probabilities of occurrence, is, in dollars,

$$E = \sum_{k=1}^{\infty} 2^{-k} 2^{k} = 1 + 1 + 1 + \ldots = \infty .$$

So if expected value were used as the index of reasonable price, a rational individual should be willing to pay a very great amount indeed to play this game. Nevertheless, it probably would surprise no one to discover that when people are asked what they would be willing to pay for the opportunity to take this gamble, almost no one would give a very large sum.

Inasmuch as an infinite amount of money is a bit beyond the reach of any finite being, we might make the situation more realistic by limiting the number of tosses allowed, say by agreeing to consider the game to be terminated either by the toss of a head or with the nth toss (irrespective of what it is), whichever came first; if the nth toss is a head, you win $\$2^n$ and if it is a tail, you win nothing. (The idea of limiting the number of tosses so as to avoid the need to deal with infinity was suggested at least as early as 1777 by Georges Louis de Buffon [Todhunter, 1865/2001], who proposed that it be limited to 29 on the grounds that 2^{29} would exceed the amount of money that could be furnished by all of France.)

Suppose, for example, that we agreed to terminate the game either by the occurrence of a head or with the third toss, whichever came first. Now imagine that you have an opportunity to play this game a very large number of times. We would expect that about half of the games would be terminated by the first toss, because it was a head; in all these cases you would win $2. About one fourth of the games would be terminated by the second toss, and on these you would win $4. The remaining one fourth of the games would be terminated with the third toss, and you would win \$8 on the one half of these that are heads. So on any given game, you would win \$2, \$4, \$8, or \$0 and your average, per-game, win—the expected value of the game—would be

$$(1/2 \times \$2) + (1/4 \times \$4) + (1/8 \times \$8) + (1/8 \times \$0) = \$3.$$

Given the opportunity to play this game many times, you probably would be happy to do so at any cost to you that is much less than \$3 per game. At any cost greater than \$3 per game, you probably would not find it very attractive, because you would be almost sure to lose money in the long run. You might even be willing to pay something close to \$3 to play this game even if you could only play it once, because you have three chances in eight of winning \$4 or more, and you might consider that to be a reasonable gamble.

Now suppose that we agreed that the game would be terminated either by the toss of a head or by the 1,000th toss, whichever came first. The situation is completely analogous to the simpler one, except that the expected value of the game, in dollars, is

$$E = \sum_{k=1}^{1000} \left(\frac{1}{2^k}\right) 2^k = 1{,}000.$$

Theoretically, one should consider any cost of playing this game that is much less than \$1,000, say \$800, to be attractive. However, it seems unlikely that many people would be willing to pay more than a relatively few dollars to play it. The probability is .5 that one would win only \$2, and the probability that one would win as much as \$1,000 is only about .001. (Even in the St. Petersburg game, the probability is .5 that one will win only \$2, and, as Samuelson [1977] points out, the expected total number of coin tosses, *n*, per game is only 2:

$$E(n) = \sum_{i=1}^{\infty} 2^{-i} i = 2.$$

It is possible to win an extremely large amount of money in the game described, but the probability of doing so is very small. If one agreed to pay any sizeable fraction of the expected value of the game—for example, one half of it, or \$500—one would be very likely to lose most of what one paid. And if one paid this amount many times in order to play the game repeatedly, one would be likely to accumulate a very sizeable loss. In theory, if one could play the game infinitely many times, sooner or later a game would last long enough—the string of successive tails would get long enough—to give one a win large enough to more than offset an accumulated loss. But the time that one might have to wait before getting a win large enough to offset a cost of a size that is very likely to accumulate could be very long indeed. In fact the chance of ever getting ahead in this game in finite time is vanishingly small.

The number of dollars that the gamble *permits* to be won in a single game, 2^{1000}, or about 10^{300}, is larger by roughly 220 orders of magnitude than the number of particles in the universe according to Eddington's famous estimate. Of course, the probability that this amount will be won, 2^{-1000}, is sufficiently close to zero as to be indistinguishable from it by any measurement technique that has been, or is ever likely to be, developed.

This all illustrates a point made by Allais (1979/1990) regarding the inapplicability of the law of large numbers to small samples. *"If I am to participate in a long series of games, but could be ruined early on, possibly even in the first round, it is obvious that the justification of the rule of mathematical expectation by the law of large numbers is invalid.* There would be little consolation for me in the knowledge that, had I been able to hold on, my winnings would probably have tended to the value of their mathematical expectation" (p. 116, emphasis in the original).

The original St. Petersburg gamble involves an event that, by definition, has an infinity of possible outcomes. The variations on the gamble considered previously have a finite number of outcomes, but as in the case of the original paradox, the expected value of each gamble is the sum of the products of the values of the individual outcomes and their probabilities of occurrence. We could think of each of the cases considered as the sum of many individual gambles, each of which has an expected value of $1. The original paradox, for example, can be thought of as composed of an infinity of independent gambles, the first of which has the possible outcomes H and T and an expected value of $1 obtained from adding the values of the two possibilities each multiplied by its probability of occurrence: (1/2 × $2) + (1/2 × $0). The second gamble in the composite can be thought of as having the possible outcomes TH, TT. The only outcome in this gamble that pays off is the first, which pays $4 and does so with probability 1/4, so the expected value is again $1. And so on.

When one buys a ticket for a lottery in which half of the income from the ticket sales is to be used as the prize money, the expected value of a ticket, assuming the lottery is fair, can be no more than half its purchase price. The fact that such gambles are purchased in abundance suggests that the component gambles that make up the St. Petersburg paradox would be perceived by many people as reasonable buys at something close to their expected values, or more. Thus, given the chance to buy for $1 the individual gamble of, say, TTTTTTTTTH, which has an expected value of $1, because it will pay $1,024 with probability 1/1024 and $0 with probability 1,023/1,024, many people would undoubtedly take it.

Whether everyone who would gladly spend $1 for this gamble would pay $1000 for the opportunity to participate in the gamble 1,000 times is doubtful, even though doing so would very considerably increase their chances of winning at least once and would give them the possibility of winning several times. (The probability of winning ($1,024) once or more in 1,000 independent tries is about .62; the expected value is, of course, $1,000.) The deterrent is the fact that there would also be a fairly good chance (about .38) of winning nothing and consequently losing the entire $1,000 that was paid for the opportunity to play. It is easy to see, intuitively, why one might be happy to purchase one, or a few, of the individual gambles that comprise the St. Petersburg game for something close to, or even a bit more than, their expected values, while being unwilling to pay a lot for all of them combined.

The St. Petersburg paradox has generated a great deal of discussion and debate. A list of people who have written on the topic, from the 18th century on, would include many of the better known names in the history of mathematics: Buffon (1777), Condorcet (1785), Laplace (1812), Cournot (1843), Bertrand (1889); Keynes (1921), and Samuelson (1960). An account of contributions of these and other writers has been given by Samuelson (1977). Many resolutions of the paradox have been proposed and I think it safe to say that there remain differences of opinion as to what the proper resolution is, and even as to whether there is one.

According to one view, the paradox is illusory: "The paradox of the Saint Petersburg problem is that there is a paradox" (Jorland, 1987, p. 157). Jorland says that the real puzzle is why it took 224 years for someone to acknowledge that the infinite series involved is not summable so there is no expectation. He credits Feller (1936–1937) with being the first to do this and to raise and answer the question of whether there exists a fair stake for a game without expectation. Jorland acknowledges that the discussion and debate engendered by the St. Petersburg paradox has been "epistemologically very fertile. It led to the substitution of the law of large numbers for the principle of insufficient reason as the foundation of mathematical expectation and it raised the question of the objective or subjective nature of probability depending on whether it applies to single events" (p. 181). Despite Jorland's expression of surprise that the paradox was seen as a paradox for so long before Feller's work, he ends his discussion of it by claiming that none of the solutions of it that have been proposed is satisfactory.

Independently of the question of what the correct resolution of the St. Petersburg paradox is, or of whether there is one, there can be no question of the fact that the paradox has stimulated a great deal of thinking about probabilistic reasoning. Samuelson (1977), after having "defanged" the paradox with his own resolution, pays it this tribute: "When you have defanged a paradox with the texture of the St. Petersburg puzzle, the problem does not disappear or fade away into banality. As my *Historical Notes* [subsequent comments on the history of treatments of the paradox] illustrate, so many points were raised by commentators on the problem that the St. Petersburg paradox enjoys an honored corner in the memory bank of the cultured analytic mind" (p. 36).

Expected Value and Most Likely Outcome

The St. Petersburg paradox illustrates in a particularly compelling way that the *expected* value of a probabilistic outcome is not necessarily the value of the *most likely* outcome. This distinction is seen in the following situation, from Paulos (1992), which may be thought of as a modern analogue of the paradox. Consider a volatile stock that each year, with equal probability, increases in value by 60% or decreases by 40%. What is $1,000 invested in this stock likely to be worth in 100 years? Inasmuch as the average change in the stock's value is (+60 – 40)/2, or 10%, the expected value of the $1,000 investment after 100 years is $1,000 × 1.1^{100}, or $13,780,612; not a bad return. What your broker is less likely to tell you is that, given the assumption that the stock is equally prone to increase (by 60%) or decrease (by 40%) in any given year, the most probable scenario is that it will increase in 50 of the 100 years and decrease in the other 50, and its final value in this case will be $1,000 × 1.6^{50} × 0.6^{50}, or $130.

One's initial reaction to such disparate numbers for *expected* value and *most probable* outcome is likely to be that something must be wrong—that one of the calculations must have been done incorrectly or that the reasoning must have been based on some unstated faulty assumption. But where is the mistake? Is there something wrong with the idea of averaging the equally likely 60% gain and 40% loss to get a mean annual gain of 10%? Suppose you made two $1,000 investments and gained 60% on one while losing 40% on the other in the first year; your average gain per investment, (600 – 400)/2 = 100 dollars, would be 10%. But this is not analogous, one might argue, to the case in which you have a gain of 60% followed by a loss of 40% (or vice versa) on the same investment, in which case the total change is 600 – 640 = –40 (or –400 + 360 = –40), which is –4%. Does this mean that using the average of the equally likely percentage changes to calculate expected value of the imagined investment is not legitimate?

In fact it is legitimate. To get a better intuitive feel for the situation, it may help to trace the first few branches of the tree of possible histories of the investment. Table 6.1 gives, for each number of years 1 through 5, and 10, every possible outcome (combination of gains and losses) for the investment over that number of years, the number of ways each outcome can be realized, the probability of obtaining that outcome, the value of that outcome, the product of the

TABLE 6.1

Possible Outcomes of a $1,000 Investment That Is Equally Likely to Increase by 60% or Decrease by 40% per Year

Year	*Outcome*	*No.*	*Prob. (P)*	*Value (V)*	*P* × *V*	*Exp. Value*
1	1G	1	.500	1,600	800	
	1L	1	.500	600	300	
						1,100
2	2G	1	.250	2,560	640	
	1G, 1L	2	.500	960	480	
	2L	1	.250	360	90	
						1,210
3	3G	1	.125	4,096	512	
	2G, 1L	3	.375	1,536	576	
	1G, 2L	3	.375	576	216	
	3L	1	.125	216	27	
						1,331
4	4G	1	.0625	6,554	410	
	3G, 1L	4	.2500	2,458	614	
	2G, 2L	6	.3750	922	346	
	1G, 3L	4	.2500	346	86	
	4L	1	.0625	130	8	
						1,464
5	5G	1	.0313	10,486	328	
	4G, 1L	5	.1563	3,932	614	
	3G, 2L	10	.3125	1,475	461	
	2G, 3L	10	.3125	553	173	
	1G, 4L	5	.1563	207	32	
	5L	1	.0313	78	2	
						1,610
	***					
10	10G	1	.001	109,951	107	
	9G, 1L	10	.010	41,232	403	
	8G, 2L	45	.044	15,462	679	
	7G, 3L	120	.117	5,798	679	
	6G, 4L	210	.205	2,174	446	
	5G, 5L	252	.246	815	201	
	4G, 6L	210	.205	306	63	
	3G, 7L	120	.117	115	13	
	2G, 8L	45	.044	43	2	
	1G, 9L	10	.010	16	<1	
	10L	1	.001	1	<1	2,593

probability and the value of that outcome (which is the contribution of that outcome to the expected value for the year), and the (approximate) expected value for the year (the sum of these products for that year).

The expected value at the end of each of these years, obtained by adding up all the possible outcomes weighted by their probabilities of occurrence, is the same as we would get by using the standard compound-interest formula with the interest set at 10%. The breakdown shows how the various possibilities contribute to the expectation. It also gives a sense of how the most likely (modal) outcome steadily decreases. At the end of 10 years, the most likely value of the investment is $815, which represents a loss of $185, and 638 of the 1,024, or about 62%, of the equally-likely outcomes represent a loss of at least that amount. Continuation of this analysis would show the expected value growing increasingly rapidly, reaching $13,780,612 at the end of the 100th year, whereas the value of the most likely outcome continues slowly to decrease, and the probability of a net loss remains always greater than .5.

Saying that the expected value of this investment after 100 years is $13,780,612 is a little like saying that you and some celebrity had an average income last year of $20,000,000; true, perhaps, but probably not very exciting from your point of view. (I am making an assumption here about who is most likely to read this book.) The expected value is skewed greatly by virtue of low-probability outcomes of extremely high worth. Suppose that the stock were to move in the same direction, either to increase or to decrease in value, every one of the 100 years. The probability of either of these outcomes (assuming the direction of change is random) is vanishingly small—$(1/2)^{100}$ or approximately one chance in 10^{30}—but theoretically possible; the important point is that although the two outcomes are equally probable, their impacts on the expected value computation are very different in magnitude. Following 100 years of steady 60% increases, the value of the stock would be $\$1{,}000 \times 1.6^{100}$, or $\$2.58 \times 10^{23}$, many orders of magnitude more than the net worth of all the businesses on earth. Following 100 years of steady 40% declines, it would be worth $\$1{,}000 \times 0.6^{100}$, or, for practical purposes, 0. The average of these outcomes is $\$1.29 \times 10^{23}$, which is to say the positive scenario counts much more in the expectation than the negative one, because the latter is bounded below by 0.

The probability of 100 consecutive moves in the same direction is so small that neither of these possibilities has an appreciable effect on expected value; however, the principle that the comparison illustrates pertains to other possibilities of higher probability and thus of greater impact on the expectation. The probability that the number of years that the stock gains in value will exceed by a specified amount the number of years that the stock loses in value is the same as the probability that the number of losing years exceeds by that amount the number or gaining years; however the impact on expected value of an excess of gain years over loss years is much greater than the opposing impact of an equal excess of loss years over gain years. The overall effect of this asymmetry is that the expected value of the investment gives a highly distorted picture of what is *most likely* to happen.

The counterintuitiveness of the compound interest example is due, at least in part perhaps, to the easy-to-overlook nonequivalence of equal percentage gains and losses. In particular, if one failed to think about it, one might assume that an increase and decrease (or a decrease and increase) of a given percentage offset each other so one ends up where one began, but of course that is not the case. A loss of 50% offsets a gain of 100%, or conversely, it takes a gain of 100% to compensate for a loss of 50%. In our example, a loss of 40% more than offsets a gain of 60%, so the greater the *equal* number of gains and losses of these magnitudes the further behind one will fall. An investment that gains and loses 50% in alternate years will lose 25% of its existing value every 2 years, so at the end of *n* years it will be worth only $.75^{n/2}$ of its original value; $1,000 in this case becomes $237 in 10 years. To satisfy my curiosity, I calculated the losses that would exactly offset gains of specific magnitudes. The results, expressed in percentages, are shown in Table 6.2.

It does not follow from any of this that a decision to invest in a stock that is expected to behave as the volatile one just described would necessarily be irrational. Whether such an investment would make sense would depend on the investor's attitude toward risk. One might be inclined to invest in such a stock, although hopefully not one's last dollar, on the grounds that even though the

TABLE 6.2
Magnitudes of Losses That Would Offset Specified Gains, All in Percentages

Gain	*Loss*
10	09.1
20	16.7
30	23.1
40	28.6
50	33.3
60	37.5
70	41.2
80	44.4
90	47.4
100	50.0

most likely outcome is that one would experience a modest loss, there is a very good chance that one would realize a large gain. This is especially true if one is permitted to decide when to liquidate the investment.

Here is an analogue of the volatile-stock story that may also help to put the situation in perspective. Imagine that you are given the opportunity to participate in the following game. You are to put down $1 at the outset; this is your initial "investment" in the game. (If you prefer to imagine a larger investment, simply multiply all the dollar figures in what follows by whatever you wish; an investment of $1,000, for example, changes all the dollar figures by a factor of 1,000.) A coin is to be tossed 10 times. Every time the coin comes up heads, the current value of your investment will be increased by 60% and every time it comes up tails, the current value of your investment will be decreased by 40%. Would you find this an attractive game to play? The *expected value* of your investment at the end of the game is $2.59 (representing a gain of $1.59); the *most likely value* is about $0.82 (representing a loss of about 18 cents). You are more likely to lose money than to make any (the probability that you will lose *some* portion of the initial investment is about .62). The most you can lose is $1, whereas it is possible to win a much larger amount (over $100 in the most extreme case), and the chances of a relatively spectacular win are quite good; in particular the probability of at least doubling your investment is about .38 and the probability of increasing it by almost a factor of six is about .17. The probability of getting *k* or more heads in 10 tosses and the final value of the initial $1 investment for each value of *k* are given in Table 6.3.

In general, an excess of one type of outcome over the other (heads vs. tails in this game, gain years vs. loss years in the original volatile-stock scenario) is equally likely in both directions; for example, the probability of seven or more heads is the same as the probability of seven or more tails—it is about .17 in each case. However, one stands to win a great deal more with seven or more heads than one can lose with seven or more tails; in fact, given the outcome seven or more heads, the expected gain is $9.87 ($10.87 expected value including the initial $1 investment), and the minimum is $4.80, whereas given the outcome seven or more tails, the expected loss is about $0.91, and the maximum $1.

If we were to define the game as a sequence of 100 coin tosses, to correspond to the original stock scenario, the asymmetry would become much more extreme. Now the maximum possible (though extremely unlikely) gain would be $2.58 × 10^{20}, whereas the maximum possible loss would remain at $1, the value of the original investment. The most likely final value of the investment would be $0.13. As with the 10-toss game, the probability is greater than .5 that the final value would be less than $1, but the chances of realizing an enormous gain are quite good.

Table 6.4 shows the approximate probability of getting exactly *k* heads in 100 tosses and the value of the original $1 given k heads. (The gain in each case is the

TABLE 6.3

The Probability of Getting Exactly *k*, and *k* or More, Heads in 10 Tosses and the Resulting Value of $1, Given That a Toss of a Head Means a 60% Increase and One of a Tail a 40% Decrease

k	*Prob # Heads = k*	*Prob # Heads ≥ k*	*Approx Value of $1 Given k Heads*
0	.00098	1.000	<.01
1	.00977	.999	.02
2	.04395	.989	.04
3	.11719	.945	.11
4	.20508	.828	.31
5	.24609	.623	.82
6	.20508	.377	2.17
7	.11719	.719	5.80
8	.04395	.055	15.46
9	.00977	.011	41.23
10	.00098	.001	109.95

latter value minus $1 representing the original investment.) The gain associated with *k* heads is the *minimum* gain, given k *or more* heads; remember that the maximum loss is *never* greater than $1, in the situation we are considering.

Before turning to the table, it may be useful to recall that the formula for computing the probability of exactly H heads in 100 tosses is

$$p(H) = \frac{\binom{100}{H}}{2^{100}}.$$

Where $\binom{100}{H}$ represents the number of combinations of 100 things taken H at a time and the general formula for computing $\binom{n}{r}$ the number of n things taken r at a time is $\frac{n!}{r!(n-r)!}$. For large n, it is convenient to use Stirling's approximation for the factorial, $n! = e^{-n}n^{n}(2\pi n)^{\frac{1}{2}}$, approximately. Alternatively, one can use logarithms of factorials, often provided in books of mathematical tables at

TABLE 6.4

The Probability of Getting Exactly *k*, and *k* or More, Heads in 100 Tosses and the Approximate Value of $1 Resulting From *k* Heads, Given That a Toss of a Head Means a 60% Increase and One of a Tail a 40% Decrease

k	*Prob #* *Heads = k*	*Prob #* *Heads ≥ k*	*Value of $1* *Given k Heads*
50	.07959	.540	.13
51	.07802	.460	.35
52	.07352	.382	.92
53	.06659	.309	2.46
54	.05796	.242	6.57
55	.04847	.184	17.52
56	.03895	.136	46.71
57	.03007	.097	124.55
58	.02229	.067	332.13
59	.01587	.044	885.69
60	.01084	.028	2,361.83
61	.00711	.018	6,298.22
62	.00447	.011	16,795.25
63	.00270	.006	44,787.33
64	.00156	.003	119,432.88
65	.00086	.002	318,487.68
66	.00046	<.001	849,300.48
67	.00023	<.001	2,264,801.27
68	.00011	<.001	6,039,470.06
69	.00005	<.001	16,105,253.50
70	.00002	<.001	42,947,342.67

least for factorials up to 100! (The probabilities in the second column of the table were calculated with logarithms; the use of Stirling's approximation would have yielded very slightly larger values.)

The probability that 100 tosses will produce *at least* H heads is

$$p(k \geq H) = \frac{\sum_{k=H}^{100} \binom{100}{k}}{2^{100}},$$

and the value of the original $1 investment at the end of a game, given the occurrence of k heads, is

$$V_k = (1.6)^k(.6)^{100-k}.$$

More generally, the final value of an original investment, or stake, of an amount S, given the occurrence of k heads, $V_{s,k}$, is

$$V_{s,k} = S(1.6)^k(.6)^{100-k}.$$

The gain (or loss) is, of course, simply $V_{s,k} - S$.

Inspection of this table and a comparison of it with the table representing the 10-toss case should help provide a reasonably good intuitive grasp of the implications of participating in a game of the sort imagined. Consider first the 100-toss case. The expected value of the dollar investment at the end of the game is $13,780.61; the most likely final value is $0.13. The probability that the final value will actually by $0.13 is about .08; the probability that it will be $0.13 *or less* is about .54 (the same as the probability that it will be $0.13 or more). The probability that the final value will be as great or greater than the expected value of $13,780.61 is about .01. The probability that it will be less than $1—the probability that one will lose some amount of one's investment—is about .69. However, the probability is about .24 that the value will be greater than $6, almost .1 that it will exceed $124, and almost 1 chance in a thousand that it will be more than $800,000.

Obviously, if I were trying to convince you not to play this game, I would focus on the fact that you are much more likely to lose than to win—the odds against your winning anything are better than 2-to-1—and that the most likely outcome will leave you with only 13% or your original investment. If I were trying to convince you to play, I would emphasize the fact that the most you can lose is a dollar, and for the risk of that you are buying a very good chance of multiplying your investment several-fold (about one chance in four of at least a 500% gain), and a nontrivial chance of realizing a really spectacular windfall (one chance in a hundred of growing your investment by at least a factor of 6,000).

Let us now compare the 100-toss and 10-toss versions of the game. Essentially the arguments made about the possible outcomes of the 100-toss game apply qualitatively to the 10-toss game as well, but the differences in probabilities are instructive. Perhaps the most obvious difference in the two probability distributions is the fact that an outcome that deviates from the most likely one (one half heads) by a given percentage is much greater in the 10-toss than in the 100-toss game. The probability of getting 60% or more heads is about .38 in the 10-toss game and less than .03 in the 100-toss one; the probability of getting

70% or more heads is about .17 in the former case and only about 8 chances in 100,000 in the latter. The directions of these differences are not unexpected, of course; they illustrate one implication of the well-known law of large numbers according to which deviance of a given percentage is less likely in large random samples than in small ones, but the magnitude of the differences may take even some people who are familiar with the law by surprise. The disparity in the likelihoods of outcomes that differ from the modal one by a given percentage is reflected in the payoffs associated with such outcomes. In the 10- and 100-toss games, the minimum final values of a $1 investment given 60% or more heads are $2.17 and $2,361.83 respectively; for 70% or more heads, the two respective numbers are $5.80 and over $42 million.

Now, if offered a chance to play either of these games, what is the rational thing to do? In my view the answer is, it depends on one's attitude toward risk. One might find both games attractive and be delighted to play either of them, if given the opportunity, the rationale being that $1 is a small amount to risk for the relatively good chance of winning much more than that. One might be less enthusiastic if the minimum initial investment allowed were $1,000, because one is much less sanguine about the prospect of losing this amount of money even though the potential winnings are commensurately larger. If given the opportunity of playing either game as often as one wished, and terminating play when one wished, one would find it attractive even for quite large stakes, because the likelihood of eventually winning more than enough to offset a series of losses is very good.

Now suppose you were given the opportunity to play one of these games for a specified number of times. The rules in this case are that in order to play at all, you must play the specified number of times, no more and no less. The situation becomes more interesting. Suppose, you were offered the opportunity to play the 10-toss game precisely M times, but, in addition to the $1 investment you were required to make in each game, you had to pay a one-time participation fee. How much, if anything, would you—how much should you—be willing to pay? I do not propose to work out the details of this situation, but it will be interesting to do a partial analysis of a few particular cases. Consider the case of $M = 5$. How much would you be willing to pay to play the 10-toss game five times?

It is easy enough to calculate the expected value of your total ($5) investment at the end of the five games; it is simply five times the expected value of the $1 investment at the end of a single game, or $12.95; this represents an expected gain of $7.95. But as I hope the foregoing discussion has made clear, this does not tell us all that we would like to know about what the more probable of the possible outcomes are. One thing we might do is calculate the probability of getting at least one win of a given size in a series of five games. Table 6.5 gives these probabilities. (Excepting the case of $k = 0$, the probabilities

TABLE 6.5

Probability of Getting *k* or More Heads at Least Once in Five 10-Toss Games and Approximate Value of $1, Given *k* Heads

k	*Prob # Heads ≥ k at least Once in Five 10-Toss Games*	*Approx Value of $1 Given k Heads*
0	1.000 (approx)	<.01
1	1.000 (approx)	.02
2	1.000 (approx)	.04
3	1.000 (approx)	.11
4	1.000 (approx)	.31
5	.992	.82
6	.906	2.17
7	.611	5.80
8	.245	15.46
9	.053	41.23
10	.005	109.95

given as 1 are not exactly 1, of course, but round off to 1 when carried to 5 places beyond the decimal point.)

Thus the probability of having at least one out of five games increase the value of the initial investment by at least a factor of 5.8 is about .61 and the probability of having at least one game increase the value by at least a factor of 15.46 is about .25. A similar analysis for some of the possible outcomes of five 100-toss games is given in Table 6. 6.

So the probability of having at least one game produce 55 or more heads and, consequently, at least a 17-to-1 return on the investment is about .64. The probability that at least one game will produce at least 60 heads and therefore a 2,361-to-1 return is about .13.

It should be clear that these games become increasingly attractive the larger the number of times one is permitted to play them. If one is permitted to play the game 100 times, the probabilities comparable to those in the preceding table are as as shown in Table 6.7.

In other words, the probability of having at least one game produce a 2,362-to-1 return is about .94, and the probability of having at least one produce a return of over 300,000-to-1 is a non-neglible .16. These figures might make many people willing to pay a fair amount of money for the opportunity to play the 100-toss game 100 times.

TABLE 6.6

Probability of Getting *k* or More Heads at Least Once in Five 100-Toss Games and Approximate Value of $1, Given *k* Heads

k	*Prob # Heads ≥ k at Least Once in Five 100-Toss Games*	*Value of $1 Given k Heads*
55	.638	17.52
60	.134	2,361.83
65	.009	318,387.68
70	.000	42,947,342.67

TABLE 6.7

Probability of Getting *k* or More Heads at Least Once in 100 100-Toss Game and Approximate Value of $1, Given *k* Heads

k	*Prob # Heads ≥ k at Least Once in 100 100-Toss Games*	*Value of $1 Given k Heads*
55	1.000(approx)	17.52
60	.944	2,361.83
65	.164	318,387.68
70	.008	42,947,342.67

Comparison of the 1-game, 5-game, and 100-game scenarios illustrates that the expected value of the outcome becomes increasingly meaningful the greater the number of times the situation is to be encountered. When a gamble is involved, one may do well to attach significance to the expected value if one is permitted to make the bet many times, but one may wish to discount it considerably, and give more weight to other considerations, like the distribution of more likely outcomes, if one is permitted to make it only once or a few times.

In general, as Keynes (1921) and others have pointed out, expected value is a more complex concept, psychologically, than its use in computations might suggest. Mathematically, the expected value of a good worth $10,000,000 and probability of .001 is $1,000 more than that of a good worth $10,000 and probability of .9. Many people would prefer the latter gamble to the former, however, and who is to say that this would be an irrational choice?

Value and Utility

The St. Petersburg paradox generated a great deal of interest among early probability theorists and much debate about what would constitute an acceptable resolution of it. Of particular importance for the subsequent development of theories of decision making and rational behavior is the role it played in forcing a distinction between monetary value and "utility," a distinction promoted by Daniel Bernoulli (1738)—he used the terms "physical fortune" and "moral fortune"—in his original attempt to resolve the paradox. Samuelson (1977) credits Cramer (1728) as the originator of the idea of diminishing marginal utility in a letter to Nicholas Bernoulli, and published by Daniel Bernoulli in his 1738 paper. According to this idea, the utility to an individual of an increase of a given amount in her fortune will vary inversely with the amount of wealth she already has. Many of the proposed resolutions of the paradox were based on this idea—which was often expressed as a concave relationship between wealth and utility—and, in particular, on the assumption that increase in utility diminishes essentially to zero as wealth increases indefinitely.

The distinction between value and utility was also used extensively by von Neumann and Morgenstern (1953/1944) in their seminal work on decision theory in its modern form. This distinction recognizes that what a specified amount of money is worth to individuals depends on how much they have; to a person who has only $1,000, the possibility of losing $1,000 presumably is much more than 1,000 times worse than the possibility of losing $1, but this probably is not the case for one whose net worth in measured in millions. Bernoulli assumed that how much satisfaction—increase in moral fortune—one derives from a given small increase in one's physical fortune is inversely proportional to the size of one's physical fortune before the increase. This principle is incorporated in expected utility theory as the assumption that utility is a concave (looking up from the abscissa) function of money (Pratt, 1964).

The relationship between money and utility is further complicated by the recognition that the subjective value of a given amount of money may depend not only on how much money one has to begin with but on whether the money in question is a (real or potential) gain or loss. To the individual who has only $1,000, the prospect of gaining another $1,000 is likely to be very attractive, but the degree of happiness caused by this eventuality is unlikely to match the degree of sadness that would accompany a loss of the same amount. In general, it appears that losses have greater subjective value than gains of equal amounts (Galanter & Pliner, 1974).

The St. Petersburg paradox and the variations on it considered earlier not only illustrate the necessity of the distinction between monetary value and utility, or between objective and subjective worth; they also demonstrate that the attractiveness of a possible venture with an uncertain outcome can depend not only on the mathematical expectation of the outcome but on the entire probability distribution of possible outcomes. A distribution that is highly skewed—with a great difference between mean and mode—represents a very different situation from one that is symmetrical about its mean, and, as Allais (1979/1990) has pointed out, an attitude that takes account of the dispersion of psychological values should not be considered irrational.

The stir that the paradox made and the rethinking that it caused of the meanings of such foundational concepts as probability and expectation also illustrate the importance of intuitive notions of reasonableness as the final court of appeals on questions of what shall pass for rationality and what shall not. Any theory of rationality that prescribed that one should be willing to pay an infinite amount of money, or even a large amount, for the opportunity to make the St. Petersburg wager would be a strange guide for behavior indeed.

Zabell (1993) has argued that probability theory is useful precisely because it does not correspond to our intuitions at all points; if it did, he says, we would not need it. The point is well taken in the sense that our immediate and casual intuitions when thinking about probabilistic events are often wrong, and a consultation of probability theory can set them right. When, however, what the theory has dictated has violated the intuitions of people who have thought deeply and long about the issues, either the basis of the faulty intuitions has become understood and the intuitions changed, or the theory has been modified to be consistent with the intuitions that persist. Intuition is the final judge.

Bertrand's Paradox

A paradox attributed to Joseph Bertrand is a case in which there appears to be more than one answer to a question of probability, depending on how one looks at a situation. One form of the problem is to state the probability that a randomly drawn chord of a circle is longer than a side of the circle's inscribed equilateral triangle. In terms of Fig. 6.1, the question is, what is the probability that a randomly drawn chord will be longer than the line *AC*. Here are three answers and their rationales.

Answer 1: 1/2. A radius drawn perpendicular to *AC* is bisected by *AC*, which is to say the length of the line from the origin to the point of intersection is *r*/2. (This is easily proved.) Any chord whose perpendicular distance to the origin is smaller than *r*/2 is longer than *AC;* any chord whose perpendicular distance to the origin is larger than *r*/2 is shorter than *AC*. The probability that a randomly drawn chord will have a perpendicular distance from the origin smaller than *r*/2 is 1/2.

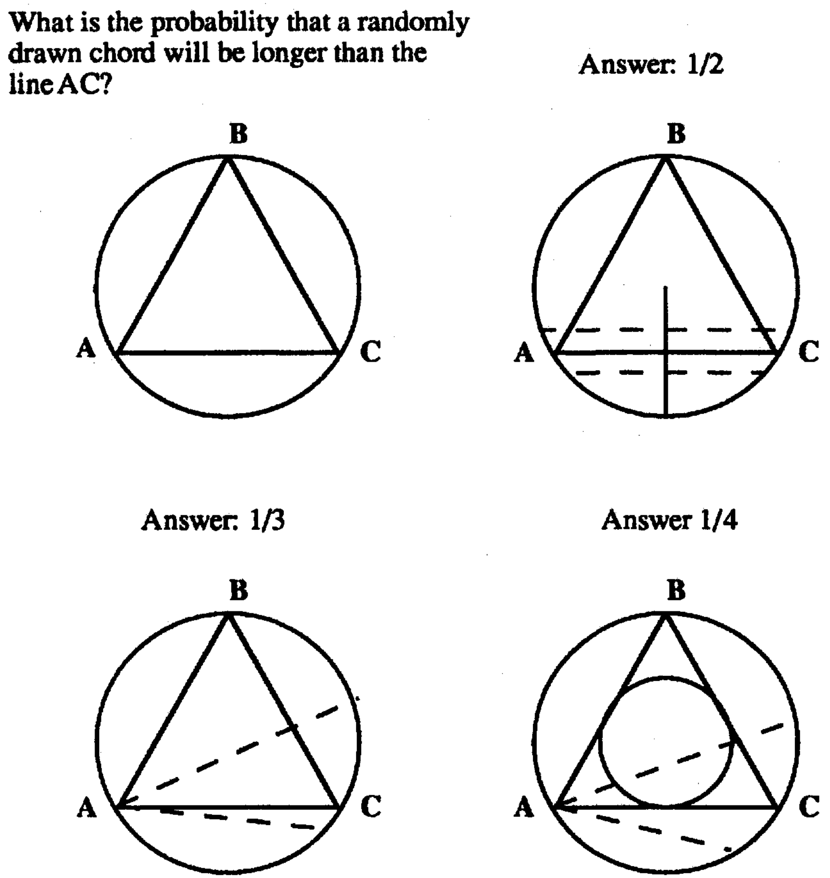

What is the probability that a randomly drawn chord will be longer than the line AC?
Answer: 1/2
B
A
C
B
A
C
Answer: 1/3
Answer 1/4
B
A
C
B
A
C

FIG. 6.1. Bertrand's paradox. Illustrating three possible answers to question of the probability of a randomly drawn chord of a circle being longer than the side of an inscribed equilateral triangle.

Answer 2: 1/3. Each end of the chord is equally likely to be anywhere on the perimeter of the circle. Suppose one end is at A. Draw an inscribed equilateral triangle with a vertex at A. The vertices of this triangle divide the circumference into three equal arcs. The chord originating at A will be longer than a side of the triangle if and only if it terminates on the arc opposite the triangle's vertex at A. Inasmuch as it is equally likely to terminate anywhere on the perimeter, the probability that it will terminate on the arc opposite A is 1/3.

Answer 3: 1/4. For a chord to be longer than AC its midpoint must lie within the circle inscribed inside the inscribed triangle. The midpoint of a randomly drawn chord is equally likely to fall anywhere within the larger circle, so the probability of it falling within the smaller circle is the ratio of the area of the smaller circle to that of the larger, which is 1/4.

What is the resolution of this paradox? Is there one? Each of the answers to the original question appears to have a legitimate rationale. Is one and only one of them correct? Can they all be correct? And if so, what does this do to our basic ideas about what probability means? If a well-formed question of probability can have two or more incompatible answers, is the theory of probability not inconsistent and therefore of dubious value?

The difficulty here is similar in principle to the basis for confusion in some of the problems discussed in the preceding sections. At the heart of it is the fact that the statement of the problem does not provide enough information to let one come up with an unqualified answer that is obviously correct. Specifically what is missing is a definition of what is meant by "randomly drawing" a chord, and this is critical because the concept is ambiguous.

Imagine that we decide to do an experiment in order to determine empirically which, if any, of the aforementioned answers to the Bertrand paradox is correct. We propose to draw at random a very large number of chords and see what percentage of them are longer than *AC*. This is a brute-force way to settle the issue. But it turns out not to be that simple, because without first settling the question of what it means to draw a random chord, we cannot proceed.

Suppose you were given the task of specifying the chords. How would you go about guaranteeing that each chord is a random selection? Having just thought about Bertrand's paradox, you might consider several possibilities. Here are three rules that might come to mind.

Rule 1. Draw a number between 0 and *r* from a table of random numbers to determine the perpendicular distance, *d*, of the chord from the center of the circle. Draw a second number between 0 and 360 to determine the orientation of the chord with respect to the perimeter of the circle.

Rule 2. Draw a number between 0 and 360 from a table of random numbers to determine the location of one end of the chord. Draw a second number between 0 and 360 to determine the location of the other end of the chord.

Rule 3. Draw a pair of numbers from a random number table, each between -*r* and *r*. Use the two numbers of each pair to represent the *x* and *y* coordinates of a point on a square Cartesian grid superimposed on the circle. (Some of the points—those near the corners of the square—will fall outside the circle; ignore these points.) When a point falls within the circle, use it as the midpoint of a chord.

It should be clear that if Rule 1 is used, the experiment will yield Answer 1 and that Rules 2 and 3 will give Answers 2 and 3, respectively. Other rules could be stated that would yield different results. None of these rules is correct in an absolute sense; they represent different operational definitions of what it could mean to select a chord at random. Without the degree of specificity expressed in these rules, the idea of selecting a chord at random is not sufficiently precise to admit of only one interpretation. In short, if asked what odds you would accept on a bet that a randomly drawn chord will be longer than *AC*, before answering you should ask what random process is to be used to determine the chord that is to be drawn.

To be useful, the answer must describe a procedure. It is not enough to specify, for example, that a point is to be selected at random to serve as the midpoint of the chord. The selection process represented by Rule 3—which is based on the use of Cartesian coordinates—will yield a chord longer than *AC* with probability 1/4. One could just as well select midpoints by using polar coordinates, selecting a rho value between 0 and r and a theta value between 0 and 360, and in this case, the probability that the midpoint would lie within the inner circle—and thus the probability that the chord would be longer than *AC*—would be 1/2.

Is one of the possible interpretations more natural than the others? Can we conceptualize a physical experiment that would make these kinds of qualifications unnecessary? What if we dropped rods, greater in length than the diameter of the circle, onto the circle from some distance above it and noted the chords that were formed by those rods that landed in such a way as to form them? This does not really help because we would have to be specific about the precise conditions under which the rods were to be dropped: where the center of the rods would be relative to the center of the circle, what orientation they would have, and so on.

One interpretation of what constitutes a randomly drawn chord is a chord selected from all possible chords in such a way that every chord in the set of possibilities has an equal chance of being the selected one. From this perspective, what we need to do to resolve Bertrand's paradox is determine what proportion of *all possible* chords are longer than *AC*. But how are we to do this? We cannot simply draw all possible chords and count them, because there are infinitely many. (There are also infinitely many chords that are longer than *AC*, and infinitely many that are shorter as well.)

To get around this problem, we could decide to quantize the space and consider only some countable subset of all the possibilities, making the assumption that the lengths of the chords in our subset will be representative of the lengths of the chords in the total population. (This is a big assumption, but let us make it, at least for the moment.) Now suppose we quantize the space by dividing r, the radius of the circle into 100 units of equal length, and let a chord with a given perpendicular distance from 0 have any one of 360 orientations with respect to the perimeter of the circle. (These numbers are arbitrary and their values do not affect the argument; I use 360 orientations simply because we conventionally divide circles into 360 degrees.) Given this quantization, any measurement along r must be expressed in terms of an integral number of $r/100$ and angular measures must be to the nearest degree.

With these constraints, a specific chord can be at any of 100 perpendicular distances from 0 and it can have any of 360 orientations with respect to the perimeter, so the total number of possible chords in the quantized space is 36,000. It should be clear that precisely half of these, those for which the perpendicular distance from 0 is less than $.5r$, are longer than *AC*. So we appear to have here a vote in favor of Answer 1 as the correct one. But, in fact, there are more ways than one to quantize the problem space, and the one we have just considered happens to be a variant of Rule 1.

An alternative way to quantize the space is to divide the perimeter of the circle into 360 units of equal length, as before, and consider the total population of chords to be all the chords that can be drawn between all possible pairs of perimeter points. There are $(360 \times 359)/2 = 64{,}620$ such chords. For a chord to be longer than *AC* it must be drawn between perimeter points that differ by more than 120 and less than 240 units. The number of chords in this subset is $(360 \times 120)/2 = 21{,}600$ or approximately one third of the total 64,620. This seems to be a vote in favor of Answer 2. But again, this is because our quantization scheme is a variant of Rule 2.

It will come as no surprise that we can base our quantization on a variant of Rule 3 and get a count that favors Answer 3. In this case we divide the space in terms of orthogonal equally spaced coordinates—we impose upon it a Cartesian grid—and let each point of intersection of two coordinates be the center point of a chord. (Assume that the grid is sufficiently fine that there are lots of points in the space.) For a chord to be longer than *AC*, its center must lie within a circle of radius $r/2$ and it is clear from the way we have quantized the space that about one fourth of all the points in it do.

So the decision to determine what percentage of all the chords that could be drawn in a quantized space have a length greater than *AC* does not give us an unequivocal answer to Bertrand's problem, because the answer we get depends on how we decide to quantize the space. The point is, Bertrand's paradox is a paradox because of the imprecision of language, or more specifically, because "randomly drawn chord" is an ambiguous term. Once one is specific about which of several possible meanings is intended, there no longer is a paradox. There is then one and only one correct answer, and experiment will bear it out. Change the meaning of the term and the answer changes as well; there is now a different answer, but there is still one and only one that is correct.

The problem of determining the probability that a randomly drawn chord of a circle is longer than the side of the circle's inscribed equilateral triangle is only one of the ways in which Bertrand's paradox has been exemplified. Another problem that is sometimes used to illustrate the paradox is the following one, from Salmon (1974):

> Suppose a car has traversed a distance of 1 mile, and we know that the time taken was between one and two minutes, but we know nothing further about it. Applying the principle of indifference, we conclude that there is a probability of 1/2 that the time taken was in the range of 1 to 1 1/2 minutes, and a probability of 1/2 that the time taken was in the range 1 1/2 to 2 minutes. A logically equivalent way of expressing our knowledge is to say that the car covered the distance at an average speed between 30 and 60 miles per hour. Applying the principle of indifference again, we conclude that there is a probability of 1/2 that the average speed was between 30 and 45 miles per hour, and a probability of 1/2 that the average speed was between 45 and 60 miles per hour. Unfortunately, we have just been guilty of self-contradiction. A time of 1 1/2 minutes for a distance of one mile is an average speed of 40, not 45, miles per hour. On the basis of the same information, formulated in different but equivalent terms, we get the result that there is a probability of 1/2 that the average speed is between 30 and 40 miles per hour, and also that there is a probability of 1/2 that the average speed is between 30 and 45 miles per hour. Since it is not impossible that the average speed is between 40 and 45 miles per hour, the foregoing results are mutually incompatible. (p. 94)

Is this a paradox or a faulty analysis? One might argue that it is the latter and that the fault lies in the assertion that the two ways described in the problem statement of expressing our knowledge of the situation are "logically equivalent." Assuming that there is a probability of 1/2 that the time taken was in the range of 1 to 1½ minutes is *not* logically equivalent to assuming that there is a probability of 1/2 that the average speed was between 30 and 45 miles per hour. More generally, assuming the equiprobability of all possible times between 1 and 2 minutes is not equivalent to assuming the equiprobability of all possible speeds between 30 and 60 miles per hour, because the relationship between time taken and speed is curvilinear.

But now, given that all one knows about the situation is that a specific distance has been traversed within some time between specified limits, what is the appropriate assumption to make? The equiprobability of all possible times? The equiprobability of all possible speeds? Something else, say a "splitting of the difference" between these two? Unless one has some reason for preferring one assumption over the other possibilities, it is not clear this question has any correct answer. The important point, for present purposes, is that and *if* one assumes equiprobability of all possible times, one gets one answer, *if* one assumes equiprobability of all possible speeds, one gets a different one, and there is nothing mysterious or paradoxical about this; it follows simply from the relationship between speed, time, and distance traversed.

Intransitivity Paradoxes

Psychologists have devoted a considerable amount of experimentation to the determination of how people solve transfer-of-inference or linear-syllogism problems of the following sort: Cheryl is taller than Pat, Phyllis is shorter than Rose, and Pat is taller than Rose; who is the tallest? One question of interest concerns the extent to which people make use of visual images or mental models in solving such problems.

In the case of this example, the order of the individuals named from tallest to shortest is Cheryl, Pat, Rose, and Phyllis. Given that we know that Cheryl is taller than Pat and that Pat is taller than Rose, we can infer with certainty that Cheryl is taller than Rose. Transitivity of ordering relationships is the rule in deterministic contexts. If A costs more than B and B costs more than C, then A costs more than C; if X is heavier than Y and Y is heavier than Z then X is heavier than Z.

Of course, any claim of having stated a rule is an invitation to find a counterexample to it, which would show it not to be a rule after all. With respect to the claim that transitivity of ordering relationships is the rule in deterministic contexts, one might point out that the United States is to the west of Spain, Spain is to the west of Japan, and Japan is to the west of the United States. This seems to be a counterexample to our rule. Or if we imagine Tom, Dick, and Jane holding hands in a circle, we could imagine Tom being to the left of Dick who is to the left of Jane who is to the left of Tom. These counterexamples to our rule make it clear that the rule, as stated, is not quite precise enough. What I should have said is that transitivity of ordering relationships is the rule with deterministic *unidimensional* variables.

What constitutes a unidimensional variable is not always obvious. Age, weight, and height would probably be considered unidimensional variables by most people, but what about intelligence, attractiveness, or leadership. If told that pairwise comparisons had shown Sam to be taller than Pete, Pete to be taller than Joe, and Joe to be taller than Sam, most of us would probably protest that at least one of the measurements must have been wrong. But if told that a poll involving pairwise comparisons had shown Sam to be a better leader than Pete, Pete to be a better leader than Joe, and Joe to be a better leader than Sam, we might be less certain that the poll was done poorly and somewhat inclined to wonder if there is something about the complicated concept of leadership that would make such an outcome possible.

In fact, in the case of polling or voting situations, ordering intransitivities can be obtained even apart from the assumption that what is being ordered is a multidimensional variable. This has been known at least since Condorcet (1785), who pointed out that if, in a three-way contest, one third of the voters prefer A to B and B to C, one third prefer B to C and C to A, and one third prefer C to A and A to B, then a majority prefers A to B, a majority prefers B to C, and a majority prefers C to A. The possibility of such an outcome has been used to support arguments against the use of voting schemes that allow voters only to identify their first choice. But finding schemes that preclude such intransitivities or other surprising outcomes has proved to be very difficult. For a three-way contest, we might let each voter rank all three candidates in order of preference. It is possible with this scheme, as with the one considered previously, to obtain an outcome in which A is preferred to B, B to C, and C to A. This would be the case if each of the following orderings—A, B, C; B, C, A; and C, A, B—was preferred by one third of the voters.

Intransitivities of this and other sorts are often encountered in voting situations (Arrow, 1963; Black, 1958; Brams, 1976; Brams & Fishburn, 1983). Arrow has shown that determining group preferences is a very tricky business and that different voting structures, all of which have some plausible claim to being democratic, can produce very different outcomes. In particular, he proved that when individuals have preferences and these preferences are transitive, there is no way to design a voting system that will simultaneously satisfy all of a small set of properties that are generally recognized to be desirable for such a voting system to have. As Cole (1997) observes, "It is easy to show ... that election results depend directly on the choice of voting system. Even when the preferences of the voters don't change, they can choose different winners if they change the details of the way they vote" (p. 101). Barrow (1998) concludes from a consideration of intransitivity paradoxes that "a concept of rationality based on trasitivity cannot be transferred from individuals to collections of individuals by means of any reasonable rule for taking majority decisions" (p. 241), and further, that "there is no reliable way of establishing rational collective choices" (p. 247).

One might suspect, upon learning of the possibility of such a strange outcome as the one just described, that it has something to do with the fact that the intransitive ordering does not reflect the preferences of a single person but is an amalgam of many individual orderings. Surely, the preferences of a single person must be transitive. If I say I prefer Sam to Pete and Pete to Joe, it must be that I prefer Sam to Joe. But is it clear that that is so? Suppose I am trying to decide which of three houses to buy, and I intend to base my decision on three equally important factors: price, size, and location. Imagine that three houses—X, Y, and Z—are for sale, and, as it happens, my (transitive) preference ordering is X, Y, Z with respect to price, Y, Z, X with respect to size, and Z, X, Y with respect to location. At this point I will be unable to make a selection among the houses, because each of them has placed first with respect to one factor, second with respect to another, and third with respect to still another; it appears to be a three-way tie. If I were to discover that any one of the houses is no longer available, the selection between the remaining two would be easy, because one member of any pair would be preferred with respect to two factors and the other with respect to only one. But strangely, given a choice between X and Y, I would prefer X; between Y and Z, I would prefer Y; and between X and Z, I would prefer Z. Like an Escher staircase, X > Y > Z > X.

So an intransitive preference ordering situation is possible not only when the preferences of different people are combined but also when only those of a single person are involved. In both of the examples considered, however, the final preference ordering was obtained by combining other orderings; the difference was that the component orderings were obtained from different groups of voters in the first case and from the same individual in the second. Both cases may be viewed as voting situations, with all the votes coming from the same person in the second instance. We could still entertain the hypothesis that an intransitive ordering can be obtained only as a result of combining individual orderings.

This hypothesis is shown to be false by the following probabilistic situation, described first by the statistician Bradley Efron (Paulos, 1990). Assume we have four dice, A, B, C, and D, of the following descriptions. A has 4 on four faces and 0 on two; B has 3 on all faces; C has 2 on four faces and 6 on two; D has 5 on three faces and 1 on the other three. Defining the winning die as the one showing the higher number on a toss, it is easy to show that if we play an extended game with each possible *pair* of these dice, A will win over B, B will win over C, C will win over D, and in each case the winning member of the pair will beat the other member two times out of three, on average. It would seem to follow from the fact that A > B > C > D that A would win over C and D and that B would win over D. In fact C will win over A five times out of nine, B and D will break even, and—here is the most counterintuitive result—D will win over A two times out of three.

Just to round out the picture, if all four dice are rolled at once, C and D will each win one time in three, A will win two times in nine and B one time in nine Considering all possible three-way combinations, the outcomes will be as shown in Table 6.8.

So A is the winner among A, B, and C; D is the winner among A, B, and D; C is the winner among A, C, and D; and B, C, and D are equals among themselves. The reader may find it difficult to get a crisp mental model of what is going on here, but it should be clear that the uncritical application of the

TABLE 6.8

The Probability That Each Specified Die, Marked as Indicated in the Text, Will Win When Rolled in the Indicated Three-Dice Combination

	Probability of Winning			
Combination	*A*	*B*	*C*	*D*
A B C	4/9	2/9	3/9	—
A B D	2/6	1/6	—	3/6
A C D	2/9	—	4/9	3/9
B C D	—	1/3	1/3	1/3

transitivity principle that works so well in many deterministic orderings of unidimensional variables is not appropriate, although the temptation to apply it may be great.

A simpler example of intransive dice has been described by Stewart (1997). Imagine three dice, A, B, and C with faces as follows: (A) 3, 3, 4, 4, 8, 8; (B) 1, 1, 5, 5, 9, 9; (C) 2, 2, 6, 6, 7, 7. Tossing any two of these dice will yield one of nine equally likely combinations. It is easy to see, by considering all possible combinations, that when A and B are tossed B beats A (has a higher number) five times out of nine, when B and C are tossed C wins five times out of nine, and when A and C are tossed A wins five times out of nine. That is A > C > B > A.

Several related "nontransitivity paradoxes" are known (Blyth, 1972c; Gardner, 1974, 1976; P. S. Savage, 1994; Steinhaus & Trybula, 1959). A very readable account of some of them can be found in P. Hoffman (1988). One such, involving the occurrence of triplets in coin tosses, was described in chapter 5. The correspondence between this situation and Efron's paradox is seen if we let TTH > THH indicate that TTH dominates THH in the sense that if Jack chooses THH, Jill's chance of winning is guaranteed to be better than Jack's by the selection of TTH. From Table 5.1 it can be seen that, as we have already noted,

HHT > HTT > TTH > THH > HHT.

Simpson's Paradox

A paradox that has received some attention both from statisticians and from psychologists is sometimes referred to as the *reversal paradox* and sometimes as *Simpson's paradox* (Blyth 1972a, 1972b; M. R. Cohen & Nagel, 1934;

Lindley & Novick, 1981; Simpson, 1951). In general terms the paradox consists in the fact that two variables that are related in a certain way (positively or negatively) at one level of analysis may have the opposite relationship when analyzed at a different level.

The phenomenon is readily seen in analyses involving contingency tables. Table 6.9 is an example from Hintzman (1980) that illustrates how a table that is the sum of two other tables may show a relationship between the variables involved that is opposite the relationship shown in each of the component tables. The thing to note is that in the composite table ($a + b$), $P(\mathrm{X} \mid \mathrm{Y}) < P(\mathrm{X})$, whereas in both of the component tables (a and b), $P(\mathrm{X} \mid \mathrm{Y}) > P(\mathrm{X})$.

Messick and van de Geer (1981) have shown that, given a conditional probability relationship of the form

$$0 < P(A \mid B) < P(A \mid {\sim}B) < 1,$$

it is *always* possible to partition the data with respect to some third variable, C, such that

$$P(A \mid BC_i) \geq P(A \mid {\sim}BC_i),$$

for all i, which is to say that reversability is always possible: "The consequence of this result is to generalize the reversal paradox and to show that any ordinal relationship between two variables can be 'reversed' by the introduction of a third variable such that within each level of this third variable, the relationship

TABLE 6.9

The Composite Table Shows a Relationship Between X and Y That Is Opposite From the Relationship Shown in Each of the Component Tables

		Table a			*Table b*			*Table a + b*		
		X	*~X*	*Σ*	*X*	*~X*	*Σ*	*X*	*~X*	*Σ*
	Y	20	0	20	20	60	80	40	60	100
	~Y	60	20	80	0	20	20	60	40	100
	Σ	80	20	100	20	80	100	100	100	200
P(X ǀ Y)		20/20 = 1.00			20/80 = .25			40/100 = .40		
P(X)		80/100 = .80			20/100 = .20			100/200 = .50		

Note. From Hintzman (1980), Table 1. Copyright 1980 by the American Psychological Association. Adapted by permission of the author.

between the two original variables is precisely the opposite of the original relationship" (p. 588). They note too that the reversal process can be repeated (restoring the original relationship) with a further partitioning, and so on indefinitely. The implication, they suggest, is that what the relationship between two variables is taken to mean should depend on the set of variables examined in its determination.

Messick and van de Geer (1981) make essentially the same observation with respect to two-by-two contingency tables. Any such table can be decomposed into two additive tables in such a way that the direction of the relationship in the original table is reversed in both subtables, provided only that the original table contains no zero entries and has at least two observations in each of the off-diagonal cells.

That Simpson's paradox is of more than theoretical interest is illustrated by a study involving the question of gender bias in admissions to graduate programs at the University of California at Berkeley (Bickel, Hammel, & O'Connell, 1975). In 1973, of all the students who applied to the various graduate programs at Berkeley, about 44% of the men and about 35% of the women were admitted. Given a total of 12,763 applicants (8,442 men and 4,321 women), the shortfall of women is highly statistically significant and would seem to represent strong evidence of discrimination against women in the university's admissions policy.

However, when the admissions data from each of the university's graduate departments and interdepartmental graduate programs were examined in isolation, there was evidence of a bias against women in only four of them, and there was evidence of a bias against men in six. Given the results of this department-by-department analysis, it is hard to see how the university as a whole could be guilty of a bias against women in its admissions policy.

The explanation of the apparent discrepancy in the Berkeley data was that departments to which women applied in the greatest numbers tended to accept a smaller percentage of applicants than departments to which women were less likely to apply. Bickel et al. (1975) suggested that an appropriate way to check for bias on a university-wide basis would be to compare the overall number of women accepted to the number expected on the assumption of no bias where the number expected on the assumption of no bias is the sum of the number expected for each department and this number, for a given department, is the product of three terms: the percentage of people applying to that department who are women, the percentage of the people applying to that department who are accepted, and the number of people applying to the department. In other words, the expected number of women accepted by a department is taken to be the number of people applying to the department weighted by *both* the percentage of that number who are women *and* the percentage who are accepted.

How it is that evidence of a bias may appear with university-wide data when such evidence does not exist at the departmental level may be more difficult to grasp than the possibility of real bias at the departmental level being obscured in aggregate university-wide data. Intuitively it is easy to see how, if each of several departments had a bias against women and a roughly equal number had a comparable bias against men, the combined data could show no bias at the university-wide level.

The Berkeley data illustrate the possibility of coming to different conclusions regarding some aspect of behavior—in this case the behavior of an institution with respect to admissions—depending on whether one looks at the data set in the aggregate or does a finer-grained analysis. But which of the two possibilities is the correct one? E. Martin (1981) argues that "both are correct, that neither has an a priori privileged status: they are but two aspects or views of the same thing, just as latency and accuracy can be two measures made on a single process or system. Thus neither a federal agency, say, nor a university is 'lying with statistics' when it cites one or the other result in defense of some position" (p. 372).

I agree that an argument can be made for both types of analysis, but find it hard to see them as equally informative. The department-by-department analysis is more informative, in my view, than is the analysis based on university-wide data. This is seen in the fact that the results of the analysis of the university-wide data are inferable from the results of the department-by-department analysis, but the converse is not true. I would argue further that an individual who was aware of the results of both types of analysis and cited only the results of the university-wide analysis in order to support the charge of bias on the part of the university would be guilty, if not of lying with statistics, at least of being less than forthright in the use of informative data.

Hintzman (1980) argues that the possibility of Simpson's paradox makes the results of many experiments on human memory that have relied primarily on contingency-table analyses suspect, because relationships that have been revealed at one level of analysis might have been reversed had analyses been done at other levels. In response to this criticism of the way memory data have often been analyzed, E. Martin (1981) claims that there is little, if any, empirical evidence that the results of memory experiments that have been analyzed in two-by-two contingency tables are Simpson paradoxical. He does not say that they *could not* be Simpson paradoxical, but, as I understand his position, only that there is no empirical evidence that they, in fact, are.

Messick and van de Geer (1981) note that the reversal paradox can be stated either in terms of conditional probabilities or in terms of correlations between variables. In the latter case, the paradox consists in the possibility of having a positive (negative) correlation between two variables while having a negative (positive) correlation between the same varibles within each cell of a partitioning based on another variable. Messick and van de Geer point out further that, stated in this way, the paradox is related, though not identical, to the problem of illusory correlation (see chap. 9): "Such an illusory correlation can be suspected to be present when two variables are found to be correlated statistically but when the partial correlation between the variables with respect to a third variable is of less extreme value, perhaps zero, or perhaps even reversed in sign" (p. 584). They add the caution, however, that when the complete and partial correlations have opposite signs, which, if either, correlation should be considered illusory may depend on the specifics of the situation.

D. S. Wilson (1975) has shown how the reversal paradox could conceivably have played a part in the development of altruism even if, as is generally assumed, nonaltruistic individuals have a survival advantage over altruistic members of the same group. If altruists and nonaltruists are unevenly distributed over "trait groups"—relatively many altruists in some groups and relatively few in others—it could happen that altruists will have a greater survival rate than nonaltruists in the population as a whole, even though the reverse relationship holds for every trait group.

Simpson's paradox has been implicated by various investigators in other interesting phenomena, such as paradoxes of preference (see Shafir, 1993), and what has been referred to as the "hot-hand" illusion (see chap. 2). Other discussions of Simpson's paradox may be found in Good (1972), Lindley (1972), Winkler (1972), Messick and van de Geer (1981), S. H. Shapiro (1982), Wagner (1982), and Sprent (1988). Falk and Bar-Hillel (1980) make the interesting observation that a judicious redistribution of the U.S. population could increase the average IQ in all 50 states; one can imagine other contexts in which Simpson's paradox could be exploited to good cosmetic effect.

Paradoxes of Expectation

If you expect to be surprised, can you then be surprised? Is an expected surprise a surprise? Is expecting the unexpected a contradiction in terms? Numerous paradoxes have been described that relate to questions of this sort. Consider the following one that is discussed at some length by Poundstone (1990).

A judge sentences a prisoner to hang at sunrise on one of the 7 days of the following week and stipulates that the prisoner is not to know which day is to be his last—the day of the hanging is to be a surprise. The prisoner's lawyer convinces him that in concocting such a sentence, which was intended to keep the prisoner in suspense, the judge has unwittingly ensured that the hanging cannot occur at all. The argument goes as follows. The prisoner knows that the hanging cannot take place on Saturday, the last day of the week, because, if it did, it would not be a surprise—if he had survived through Friday morning, he would know that the fateful day would be the only one that was left. But, if he cannot be hung on Saturday, then the effective last possibility is Friday. Therefore, by the same reasoning as before, it cannot happen on Friday. And so on through the remaining days of the week. The prisoner, of course, is delighted by his lawyer's clever reasoning and puts his mind at ease, believing his life to have been saved by the judge's blunder. On Tuesday morning he is escorted to the gallows, to his great surprise. As Poundstone points out, if the prisoner accepts the impossibility of the order being carried out, then he can be hung on any day and it will be unexpected.

If the judge had really wanted to keep the prisoner in suspense and ensure that the passage of time would provide no information regarding the day of the hanging, he could have sentenced him to a nonaging waiting time (see chap. 2.). He could have had the jailor roll a die every morning and conduct the hanging on the first day a toss yielded a three, say. With this sentence, the probability that a given day is the prisoner's last does not change over time. He could hang on any day; and indeed the probability that he will hang today given that he did not do so yesterday is 1/6 and remains so until the end. He will hang eventually, but the expected additional waiting time remains constant at $1/p$, or 6, days, independently of how many days have already passed since the sentencing.

Paradoxes of Confirmation

A widely held view, and one that seems to be consistent with common sense, is that a case of a hypothesis is confirmatory of that hypothesis. According to this view, the hypothesis that all crows are black is supported by the observation of a black crow. This idea has been challenged. Good (19671983b) gives an example of a situation in which a case of a hypothesis can be disconfirmatory with respect to that hypothesis:

> Suppose that we know that we are in one or the other of two worlds, and the hypothesis, H, under consideration is that all the crows in our world are black. We know in advance that in one world there are a hundred black crows, no crows that are not black, and a million other birds; and that in the other world there are a thousand black crows, one white one, and a million other birds. A bird is selected equiprobably at random from all the birds in our world. It turns out to be a black crow. This is strong evidence (a Bayes-Jeffreys-Turing factor of about 10) that we are in the second world, wherein not all crows are black. Thus the observation of a black crow, in the circumstances described, uudermines the hypothesis that all crows in our world are black. (p. 119)

Gardner (1976) gives other examples of observations that are logically consistent with a hypothesis but that, nevertheless, would be likely to decrease one's confidence that the hypothesis is true. The following example he attributes to Paul Berent. Every observation of a man that is less than 100 feet tall is confirmatory with respect to the hypothesis that all men are less than 100 feet tall. But suppose a man is found who is 99 feet tall. This too is a "case of the hypothesis," but it is likely to (ought to) shake one's confidence in the truth of the hypothesis.

Gardner (1976) notes that not only can cases that are consistent with a hypothesis legitimately shake one's confidence in that hypothesis, it is possible for them to show a hypothesis definitely to be false. Imagine a deck of 10 regular playing cards containing all the values from ace to 10. Suppose the deck is shuffled and the 10 cards are placed facedown in a row. Let the hypothesis be that no card with the value *n* is in the *n*th position of the row. Now suppose that as the cards are turned over, one at a time from left to right, each of the first 9 is confirmatory of the hypothesis (does not show the value that corresponds to its position in the row) and none of them is the 10; one knows before turning over the 10th card that the hypothesis is false, which is to say that the aggregate effect of the 9 confirmatory cases is disconfirmation.

These and other paradoxes of confirmation establish that a case of a hypothesis is not necessarily confirmatory of that hypothesis. It is probably safe to assume, however, that these are the exceptions that prove the rule and that in a large majority of the real-life situations of practical interest, a case of a hypothesis can be taken as confirmatory of that hypothesis—generally speaking, observation of a black crow lends credence to the hypothesis that all crows are black.

Other Paradoxes

Many other paradoxes have been described. Some involve choices that are inconsistent with axioms of decsion theory based on the idea that rationality dictates the maximization of subjective expected utility. These include paradoxes described by Allais and Ellsberg (Dawes, 1988). Whether choices that violate such axioms are considered irrational depends, of course, on how binding one considers a particular theory—more specifically, the axioms on which it is based—to be. One must decide, in specific cases, which one finds more intuitively compelling, the axioms or the choices that violate them.

Summary

The study of probability paradoxes makes clear that probabilistic reasoning can be difficult and reveals some of the factors that can make it so. Often at the heart of a paradox is some ambiguity of terminology or some missing information, without which the statement of the problem or the description of the situation is incomplete.

Is there such a thing as a true—unresolvable—probability paradox? Or is the appearance of paradox invariably attributable to ambiguity or incompleteness in problem descriptions? Certainly some apparent paradoxes can be resolved this way. Perhaps they all can be?

DILEMMAS

One of the dictionary definitions of "dilemma" is "a problem seemingly incapable of a satisfactory solution." For some dilemmas, the "seemingly" in this definition seems an unnecessary qualification. "Sophie's choice" makes the point. For such a case, perhaps the more appropriate definition, also from the dictionary, is "a choice or a situation involving a choice between equally unsatisfactory alternatives."

Prisoner's Dilemma

Considerable attention has been given by mathematicians, economists, and psychologists to the question of what constitutes rational behavior in dealing with dilemmas, or apparent dilemmas. The prototypical situation that has been studied, in numerous variations, is the "prisoner's dilemma," one version of which goes as follows:

> Two members of a criminal gang are arrested and imprisoned. Each prisoner is in solitary confinement with no means of speaking to or exchanging messages with the other. The police admit they don't have enough evidence to convict the pair on the principal charge. They plan to sentence both to a year in prison on a lesser charge. Simultaneously, the police offer each prisoner a Faustian bargain. If he testifies against his partner, he will go free while the partner will get three years in prison on the main charge. Oh, yes there is a catch ... If *both* prisoners testify against each other, both will be sentenced to two years in jail.
>
> The prisoners are given a little time to think this over, but in no case may either learn what the other has decided until he has irrevocably made his decision. Each is informed that the other prisoner is being offered the very same deal. Each prisoner is concerned only with his own welfare—with minimizing his own prison sentence. (Poundstone, 1992, p. 118)

What makes this a dilemma is the fact that (a) each prisoner can reason that he should testify against his partner on the grounds that his own sentence will be smaller by 1 year if he testifies than if he does not, whether or not his partner testifies—which is to say the strategy of testifying dominates that of not testifying—and if one testifies and the other does not, the one who testifies will do very well (whereas the one who refuses to testify will do very poorly) indeed, but (b) if *both* prisoners testify they both will do worse than they would have if both had refused to do so. The situation may be represented by a two-by-two table showing the outcomes for both prisoners of the four possible combinations of choices of "not testifying" and "testifying" by each. The entries in each cell of the following matrix show the "payoff" to prisoner A and prisoner B, in that order, of these combinations; payoffs in this case are expressed as negative numbers, inasmuch as they represent years in a jail sentence, which each prisoner would like to minimize.

	B does not testify	*B testifies*
A does not testify	–1, –1	–3, 0
A testifies	0, –3	–2, –2

There are many variations of the prisoner's dilemma, but they all have the same basic form. In general terms, we may think of the two options available to each "player" as *cooperation* and *defection.* In all cases, the strongly preferred outcome for a given player, say A, is the one that obtains when that player, A, defects and the other, B, cooperates; however, both players prefer the outcome that occurs when both cooperate over that that occurs when both defect. The following matrix (from Poundstone, 1992, p. 120), in which the cell entries represent positive payoffs, say in dollars, illustrates the general situation.

	B cooperates	*B defects*
A cooperates	2, 2	0, 3
A defects	3, 0	1, 1

One may think of the four possible outcomes in the following terms: "There is a *reward* payoff (the $2.00 above) for mutual cooperation, which both desire more than the *punishment* payoff ($1.00 above) both receive for not cooperating. But both covet the *temptation* payoff ($3.00 above), the highly desirable outcome of a single defection, even more than the reward. Both fear being the one who doesn't defect and getting stuck with the *sucker* payoff (0 above)" (p. 120).

One explanation of why a player of this game might choose to cooperate despite the fact that an analysis indicates that he should defect independently of what he believes his opponent will do notes the possibility of confusing the situation represented by the preceding payoff matrix with that represented by the following one:

	B cooperates	*B defects*
A cooperates	2, 2	0, 0
A defects	0, 0	1, 1

In this case (which is not the prisoner's dilemma), A would reason that inasmuch as both he and his opponent will realize the greatest payoff if both cooperate and that his opponent is likely to notice this also and choose to cooperate, he should choose to cooperate. The suggestion is that A might erroneously apply this reasoning to the preceding matrix, which does represent the prisoner's dilemma. It is true in the prisoner's dilemma case that the two players will realize the greatest aggregate payoff if both cooperate, and that both are likely to realize that, but if A makes the choice on the basis of thinking only this far about the situation, he runs the risk of B defecting and leaving him holding the (empty) bag. So from a normative point of view, this would be a mistake of reasoning on A's part. What is interesting is that both A and B will be the better for it if they both make the same mistake. McCain (2000) argues that such mistakes could easily be interpreted as evidence of altruism or of higher rationality. (Of course, the fact that the choice of cooperation could be due, in some instances, to an incomplete or erroneous analysis of the situation does not rule out the possibility it could be due, in other instances, to altruism or higher rationality.)

Investigators have studied both the case in which the same participants play the same prisoner's dilemma game repeatedly an unspecified number of times, and those in which they face the situation only once. This distinction is an important one, because in the latter case, the normative prescription is always to defect, but in the former there is no guaranteed best choice and the effectiveness of any particular strategy will depend on the strategy of the other player (Nowak, May, & Sigmund, 1995). In the game of indefinite duration, one's behavior in any given instance may be determined in part by the intention of influencing the behavior of one's opponent in the future, or by that of punishing uncooperative behavior in the past. The most reliable finding from studies involving repeated play without a prespecified stopping point is that consistent cooperative (competitive) play on the part of one participant tends to evoke the same type of play on the part of the other (Axelrod, 1984).

This is a finding of some practical significance, from a social psychology point of view, because it may provide some insight into the evolution of cooperative ("reciprocally altruistic") or competitive behavior (Boyd, 1988; Trivers, 1971). A strategy that has proved to be remarkably effective in repeated-play prisoner's dilemma situations is the "tit-for-tat" strategy in which the tit-for-tat player always selects the option (cooperate or defect) that was chosen by the opponent on the preceding trial (Axelrod, 1980a, 1980b; Nowak & Sigmund, 1992; An. Rapoport & Chammah, 1965).

The original and most frequently studied form of the prisoner's dilemma involves only two players, but generalizations of it include communities in which each player opposes several others, for example, nearest neighbors (Lloyd, 1995). Axelrod (1984) has shown how mutual cooperation can emerge from a cluster of individuals committed to reciprocity in prisoner's dilemma situations even within a larger community of egoists (defectors), and how "nice" strategies (strategies based on the principle of not being the first to defect) can be stable in such communities. Hirshleifer and Rasmusen (1989) have demonstrated how the possibility of ostracizing noncooperators can allow cooperators to maintain cooperation in a repeated-play game. Cole (1997) takes the relative effectiveness of nice strategies as an argument against equating survival fitness with strength: "Just because the 'fittest' tend to survive ... doesn't necessarily mean the 'fittest' are the strongest, or meanest, or even the most reproductively profligate; the fittest may be those who learn best how to use cooperation to their own ends" (p. 121).

Simulations of extended games with various strategies—tit-for-tat, generous tit-for-tat (in which defection is occasionally responded to with cooperation), win-stay lose-shift (in which a player sticks with the preceding choice if it yielded a high payoff and switches if it yielded a low one)—have revealed the possibility of the spontaneous emergence of cooperative behavior and its persistence over time under certain conditions; transitions between predominantly cooperative behavior and large-scale defection can be relatively abrupt, however, and what causes them is not completely understood (Nowak et al., 1995).

Amnon Rapoport (1967) argues that when one can assume that one's own behavior in a prisoner's dilemma game can influence the behavior of one's opponent on subsequent trials, one may no longer have a true dilemma. This is because the set of possible outcomes is no longer completely represented by the original payoff matrix; a complete representation would have to take account of the assumed influence on the opponent's play and the effects of this on future payoffs. The dilemmatic character of the situation is most clearly represented by the case in which the participants play the game only once, so the possibility of influencing future behavior is not a complicating factor—or, in the case, of repeated play, the dilemma is most stark if we consider only the final time the participants must make a decision and they know it to be the final time.

Poundstone (1992) contends that the absolute amounts of the payoffs in a prisoner's dilemma situation are not critical, that it is necessary only that the payoffs be ranked in a certain way and that the reward be greater than the average of the temptation and sucker payoffs. The latter condition is to rule out the possibility that, in a series of prisoner dilemma games, players could win more by taking turns defecting unilaterally than by adopting the cooperative strategy in each game. Nozick (1993) challenges the idea that the relative sizes of the payoffs should make no difference to a rational participant in a prisoner's dilemma game. Contrast the following two situations (the numbers are from Nozick).

	B cooperates	*B defects*
A cooperates	*1000, 1000*	*0, 1001*
A defects	*1001, 0*	*1, 1*

	B cooperates	*B defects*
A cooperates	*3, 3*	*–200, 500*
A defects	*500, –200*	*2, 2*

These situations certainly appear to be quite different. Are they different in ways that have implications for what constitutes a rational choice? Defection dominates cooperation for each player in both cases, in the conventional sense that defection gives A (B) a larger payoff independently of what B (A) does. Are there other considerations that might lead a rational person to cooperate in either case?

Nozick (1993) argues that cooperation is the rational choice in the first situation and, more generally, that "when the cooperative solution payoffs are very much higher than the dominance ones, and when payoffs for the nonmatching actions offer only slight gains or losses over these two, then we strongly will think that cooperation is rational and will find that the dominance argument has little force" (p. 53). In the second case, A risks a big loss by cooperating (if B defects) and has the possibility of realizing a large gain by defecting (if B does not), and the payoff is very little different if they both cooperate than if they both defect. Assuming the parties are independent and have no knowledge of each other's probabilities of action, the rational thing to do in this case, Nozick contends, is to chose the dominating action (to defect).

I suspect that many people will agree that these two situations are indeed different and that reasonable people are likely to react to them differently; one can be sure, however, that not everyone who thinks about these things will see it that way. One might be unwilling to consider the choice of cooperation as necessarily rational in the first case, for example, because in selecting it one runs the risk of ending up with the sucker payoff (if the other party defects). A counterargument to this position might be that although getting stuck with the sucker payoff could be irritating, the difference between the two payoffs for one's own options, given that the other player defects, is too small to matter. (If the numbers in the first of the two preceding tables do not permit this attitude, probably some could be found that would.)

This example illustrates an aspect of the prisoner's dilemma situation that I do not think has received as much attention as it deserves. Suppose I am A and I cooperate whereas B defects, so B receives the "temptation" payoff of $1,001 and I get the "sucker" prize of $0. How I feel about this is likely to depend on my general outlook on life. If I am a mean-spirited person, and can take no pleasure in anyone's good fortune but my own—if my attitude is "If I can't have it, nobody can"—I am likely to be upset, to feel unfairly treated by B who, after all, at the cost of only $1 could have made me richer by $1,000. If, on the other hand, I am able to take some genuine pleasure in other people's good fortune, I might get some satisfaction from the fact that, given B's choice to defect, my choice to cooperate made it possible for him to gain $1,000 at a cost to me of only $1. Indeed, if I were *that* kind of person, I would probably have made the cooperative choice even if I had known in advance that B would defect.

This type of consideration is, in a sense, off limits. An assumption underlying the prisoner's dilemma, as the game is usually played, is that both players are interested only in their own individual welfare; neither has any interest in advancing that of the other. Indeed this assumption is essential to ensure that the situation be a dilemma; if the players were permitted to be interested not only in their own welfare but in that of their opponent as well, the choices, in many cases, would not be problematic. This type of consideration is very relevant to the application of the prisoner's dilemma to real-life situations; some people do take a genuine interest, at least on occasion, in the implications of their actions for the welfare of other people, even that of "opponents" in situations that one might be tempted to liken to the prisoner's dilemma. But such situations are not true prisoner's dilemmas, in the strict sense, even though they may look like them, and the various theoretical treatments of behavior in prisoner dilemma situations do not apply to them.

Nozick (1993) argues that actions can have symbolic utilities that tend to be overlooked in conventional payoff representations. In prisoner's dilemma situations, it may be, for example, that making the cooperative selection has such symbolic utility for an individual: "It may stand for his being a cooperative person in interaction with others, a willing and noncarping participant in doing ventures of mutual benefit. Cooperating in this situation then may get grouped with other activities of cooperation that are not embedded in Prisoner's Dilemma situations" (p. 56). He goes on to say:

> To say all this about symbolic utility is to say that our responses to the Prisoner's Dilemma are governed, in part, by our view of the kind of person we wish to be and the kinds of ways we wish to relate to others. What we do in a particular Prisoner's Dilemma situation will involve all this and invoke it to different degrees depending upon the precise (ratios of differences among) utility entries in the matrix and also upon the particular factual circumstances that give rise to that matrix, circumstances in which an action may come to have its own symbolic meanings, not simply because of the structure of the matrix. (p. 57)

The point is that the prisoner's dilemma is an interesting and revealing representation of a decision situation in the abstract, but that its application to real-life problems is likely to be complicated by considerations that are difficult to capture in the representation.

In the conventional prisoner's dilemma game, neither player is permitted to know the other player's choice before making his or her own. If one knew the other player's choice before making one's own, there would be no dilemma for the player in the know, and the game would not have generated the interest that it has. However, some research on the effects of knowing the opponent's choice before making one's own has revealed behavior similar in some respects to that of the student who makes one choice when either of two conditions is known to pertain but makes another choice when it is known only that one or the other of the two possibilities pertain (Shafir, 1993). Shafir and Tversky (1992) found, for example, that a quarter of the participants in their study defected when they knew their opponent's choice, independently of what that choice was, but made the cooperative response when they did not know their opponent's choice. It is as though, they suggest, these people see the situation from the perspective of individual rationality when the opponent's choice is known and from that of collective rationality when it is not.

A Related Problem

A decision problem that is similar in some respects to the prisoner's dilemma, but is different in a crucial way is illustrated by the following payoff matrix, from Dawes (1988, p. 179):

	Player B	
Player A	*Strategy 1*	*Strategy 2*
Strategy 1	0, 0	5, 0
Strategy 2	0, 5	6, 6

Although this problem is similar in structure to the prisoner's dilemma, it is not a dilemma, at least in the sense that selecting the dominating strategy differs from the cooperative one. Strategy 2 dominates for both players, and if both of them pick that strategy, they both maximize their payoff. What makes the problem interesting is the fact that many people, when put in this situation, pick Strategy 1 (McClintock & McNeel, 1966).

It appears that in this situation many people focus on the *differences* between the outcomes for the two players, and that they wish to avoid the possibility of ending up with nothing while their opponent gets 5. By selecting

Strategy 1, one guarentees that one's opponent gets nothing, and leaves open the possibility of getting 5 oneself. Five is not as good as 6, but the combination of 5 for oneself and 0 for one's opponent may be considered to be better than 6 for both. This seems a particularly ungenerous—not to say self-defeating—perspective, but not an implausible one all the same. The idea that some people are willing to pay a price in order to ensure that others do not get ahead of them is hardly unthinkable, unattractive though it may be.

Dawes (1988) points out that the selection of Strategy 1 can be made consistent with subjective expected utility theory if it is assumed that a player assigns a negative value to the other's welfare. Thus, for example, if each of the numbers in the preceding table were replaced by the difference between it and the other number in the same cell, as shown in the following matrix, then Strategy 1 becomes the dominating strategy for both players.

	Player B	
Player A	*Strategy 1*	*Strategy 2*
Strategy 1	0, 0	+5, –5
Strategy 2	–5, +5	0, 0

An alternative to the assumption that people assign negative value to the welfare of others, or to the welfare of others to the extent that it exceeds their own, is the assumption that failure to select the dominating strategy in situations like those represented by the first matrix discussed in this subsection is attributable, at least in some cases, to simple failure to understand the situation well enough to know what the dominating strategy is.

There is also the possibility that people see the situation as a competitive one, in which the objective is not to maximize one's own payoff but to do better than one's opponent. To the extent that the goal is "winning"—in the sense of amassing the most points—as it is in most game situations, then the selection of Strategy 1 is rational and the second of the two matrices discussed in this subsection is a more appropriate representation of the situation than the first. It is interesting to reflect on the question of the extent to which we are inclined, by nature or by cultural conditioning, to view situations as competitive that might more productively be seen as opportunities for cooperation.

The Ultimatum Game

Another situation that bears some resemblance to the prisoner's dilemma and the situation described in the preceding section is one that is sometimes referred to as the ultimatum game:

> Imagine that somebody offers you $100. All you have to do is agree with some other anonymous person on how to share the sum. The rules are strict. The two of you are in separate rooms and cannot exchange information. A coin toss decides which of you will propose how to share the money. Suppose that you are the proposer. You can make a single offer of how to split the sum, and the other person—the responder—can say yes or no. The responder also knows the rules and the total amount of money at stake. If her answer is yes, the deal goes ahead. If her answer is no, neither of you gets anything. In both cases, the game is over and will not be repeated. What will you do? (Sigmund, Fehr, & Nowak, 2002, p. 83)

According to some conceptions of rationality, the rational thing for the proposer to do is to offer the smallest possible amount of money, $1, and the rational thing for the responder to do is to accept that offer. In fact this not what people do when put in this situation; the modal response typically is 50% of the total, about two thirds of proposers offer between 40% and 50%; only a small percentage (usually less than 5) offers less than 20% and such offers are often rejected (Camerer & Thaler, 1995; A. E. Roth, 1995; Sigmund et al., 2002).

Although there is evidence of cultural effects on precisely how people play this game (Henrich, 2000), in none of 15 small-scale societies investigated in one study did people respond in the way that the assumption that behavior is determined exclusively by self-interest would predict (Henrich et al., 2001). Individual studies in such places as Ljubljana (Slovenia), Yogyakarta (Indonesia), Tokyo, and several U.S. cities have corroborated the finding that mean offers by proposers tend to be between 40% and 50% and low offers are frequently rejected (Henrich, 2000). An exception to this rule was found by Henrich (2000) among the Machiguenga in the Peruvian Amazon, where offers averaged 20% (the modal offer was 15%) and were almost always accepted. Another, but less extreme exception, was reported by A. E. Roth, Prasnikar, Okuno-Fujiwara, and Zamir (1991), who found that Israeli proposers offered only 36% (the modal offer was 50%) but were likely (.71) to reject offers of less than 20%. Henrich interpreted his results as a challenge to the assumption "that all humans share the same economic decision-making processes, the same sense of fairness, and/or the same taste for punishment" (p. 978).

Sigmund et al. (2002) interpret the typical finding as evidence that "real people are a crossbreed of *H. economicus* and *H. emoticus,* a complicated hybrid species that can be ruled as much by emotion as by cold logic and selfishness," and see it as a challenge "to understand how Darwinian evolution would produce creatures instilled with emotions and behaviors that do not immediately seem geared toward reaping the greatest benefit for individuals or their genes" (p. 84). They see it also as evidence that people generally take account of a coplayer's view of the situation and that they place a high value on fair outcomes.

Other accounts of what appears to be magnanimous behavior in the ultimatum game have questioned the players' understanding of the situation (e.g., the fact that the game is to be played only once, when that is the case), or proposed that the utility that some players are trying to maximize differs from the game's payoff (Nowak, Page, & Sigmund, 2000). With respect to the latter proposal, we should not overlook the possibility—high probability in my view—that some people's utility functions include a genuine interest in the well-being of others that cannot be reduced to pure self-interest.

When the same players play the game repeatedly for an unspecified number of times, and any given player is sometimes the proposer and sometimes the responder, punishment of proposers who make low offers can modify their behavior in the direction of higher offers (Ostrom, Walker, & Gardner 1992). Even if the situation is structured so that low proposers cannot be identified individually (and thereby acquire reputations), low offers may elicit punishment reactions and these can be effective in increasing subsequent offers (Fehr & Gächter, 2000). Generally, in the absence of the ability to punish free-riding in public-goods experiments of the prisoner's dilemma or ultimatum game type, cooperation deteriorates badly over time (Fehr & K. M. Schmidt, 1999); however, the credible threat of punishment for free-riding can eliminate or severely reduce the practice.

Nowak et al. (2000) have shown by computer simulation that fairness (the offer by proposers of much more than the minimum) can evolve within a community of players, even when no two players interact more than once, if proposers have a means of finding out what offers have been accepted or rejected by responders in the past, but that when they do not have this means, evolution promotes low offers and low demands. As the authors put it, "When reputation is included in the Ultimatum Game, adaptation favors fairness over reason" (p. 74).

In the aggregate the results of experimentation with the ultimatum game (and other games that involve decisions regarding cooperation or sharing) support the idea expressed as a conjecture by Fehr and Gächter (2000) that "in addition to purely selfish subjects, there is a nonnegligible number of subjects who are (i) conditionally cooperative and (ii) willing to engage in costly punishment of free-riders" (p. 984). As evidence of the first point, Fehr and Gächter cite work by Fehr, Kirchsteiger, and Riedl (1993) and Berg, Dickhaut, and McCabe (1995), and as evidence of the second point work of A. E. Roth (1995) and Fehr, Gächter, and Kirchsteiger (1997).

In applying the results of experiments of the type reviewed here to real-life decisions, we should note too the possibility that some people's behavior may be motivated, at least some of the time, by true altruism, by which I mean a genuine interest in the welfare of others, independently of how their welfare relates to one's own. There are ways, of course, to see altruism as basically selfish—people can, for example, get considerable gratification from public recognition of acts of altruism. But many people also do altruistic things anonymously, and though it is possible to see such acts as selfish also—one gets the satisfaction of self-esteem—this seems to me tantamount to ruling out unselfish acts by definition.

Social Dilemmas

The prisoner's dilemma is seen as prototypical of all situations that are characterized by the "temptation to better one's own interests in a way that would be ruinous if *everyone* did it" (Poundstone, 1992, p. 126). Sociologists and economists have identified many situations in which when members of a group act in their individual self-interest they appear to be acting against the best interests of the group. Hollis, Sugden, and Weale (excerpted in A. Fisher, 1988) characterize such situations in which enlightened self-interest is self-defeating as those in which one's order of preferences is as follows:

1 You do X, others do Y.

2 You do Y, others do Y.

3 You do X, others do X.

4 You do Y, others do X. (p. 163)

They illustrate the distinction between individual and collective self-interest by reference to public goods, which are goods that are supplied to all members of a group if they are supplied to any of them: roads, public transportation facilities, public radio broadcasts. People can avail themselves of such goods whether they have contributed to the payment for them or not, so why would anyone pay? X in this example stands for not paying and Y stands for paying.

The dilemma is that when, in such situations, everyone acts in one's self-interest, narrowly conceived, no one benefits. The practical problem for the species stems from the fact that life presents many situations in which behavior that is beneficial to the individual who engages in it is disadvantageous to the individual's community or to society in general, and may be devastating if done on a large scale. Metaphors other than the prisoner's dilemma that have been used in reference to such situations include the "social trap," which Platt (1973) likens to a trap used to catch fish that makes it easy for the fish to enter but very difficult for them to get out, and the "tragedy of the commons" (Hardin, 1968). Dawes (1988) refers to such situations as "social dilemmas," which he defines as situations in which "each individual is confronted with a choice between *dominating* and *dominated* strategies, and everyone involved prefers universal choice of the dominated strategies to universal choice of the dominating ones" (p. 190).

In Hardin's (1968) tragedy-of-the-commons metaphor, a herdsman can realize a substantial personal benefit at very little personal cost by adding an animal to his herd that is grazing on common land. The benefit that comes from having an additional animal is his alone, whereas the cost, in terms of slightly less grazing land per animal, is shared by all users of the common, so the herdsman with the additional animal realizes a net gain. But of course every herdsman sees the situation the same way, so with each person working in what appears to be his own best short-term interest, they collectively ruin the land.

This is only one of many examples that could be give in which if everyone shared the same expectations and acted on them chaos would result. Consider the stock market and imagine what would happen if every trader expected the same moves in a given stock, or the maket in general, at the same time. Everyone would want to sell (buy) at the same time and consequently there would be no one to buy (sell), so the system could not work. Another way to put this is that if one finds a system that accurately predicts movements in the market, it can be effective only so long as it is not generally known; if traders sufficient in number to control the market learned of the system and used it to guide their own behavior, their behavior would ensure that its predictions would no longer be accurate.

A social dilemma of the commons-tragedy sort is likely to become highly visible to the general public when the situation approaches the point at which the collective demand on a resource seriously threatens to exceed the supply soon, but it may remain relatively invisible before that time. What to do about social dilemmas—how to get individuals to opt for dominated strategies, to act as they would prefer that everyone acted, before they reach crisis proportions—is a major societal challenge. The problem is difficult, in part, because the immediate benefit is so much more salient to the resource consumer than is the distrubuted cost (Crowe, 1969; Latané, Williams, & Harkins, 1979; Meux, 1973). Edney (1980) points out that concern about how to deal with commons problems goes back to antiquity; he notes too, however, that such problems have grown more extensive with the passage of time and the increasing size of the world population.

Poundstone (1992) suggests that the only satisfying solution to prisoner's dilemmas—of which social dilemmas are a type—is to avoid them. Good advice, no doubt, and perhaps, as he also suggests, that is what we are trying to do with ethics, laws, and other social conventions that are designed to promote cooperation or at least protect us from our own folly. M. Olson (1965) argues that some form of coercion or inducement is essential to ensure behavior in the interest of the common good in commons-tragedy situations, because the individual would be acting irrationally to pay a cost for a benefit knowing that receiving the benefit was not contingent on doing so. Commons problems are especially challenging to democratic societies, because of the conflict they pose between protecting individual freedoms and safeguarding common assets (Heilbroner, 1974; Ophuls, 1973).

One consequence that the enactment of laws and regulations can have is that of transforming dilemmas into nondilemmas by, in effect, changing the payoff matrix associated with action alternatives. The use of fines and other sanctions can change previously dominating choices into dominated ones by attaching costs to them that they did not originally have. In the classical prisoner's dilemma, enforceable threats and contracts are prohibited by definition, whereas in many, perhaps most, real-life situations these are possibilities, and this fact is important to an understanding of how cooperative behavior can arise in situations that otherwise would constitute social traps (Axelrod, 1984; Axelrod & Hamilton, 1981).

Discussions of social dilemmas often come to the basic question of what constitutes rationality (Kahan, 1974). According to some conceptions of what it means to be rational—for example, acting always in one's personal best interest—coercion may offer the only feasible way to safeguard common assets (M. Olson, 1965). It may be possible, however, to conceive of rationality in such a way that makes behavior in the interest of the common good rational, even when such behavior exacts an avoidable personal cost (Arrow, 1963; Messick, 1973). The question of the relevance of social dilemmas to our conception of rationality is an especially interesting one, because it raises reflection to a higher level of abstraction by forcing consideration of what, in view of such dilemmas, constitutes a rational conception of rationality.

Edney (1980) suggests that a fruitful way of looking at the commons dilemma is as a conflict of human values, rather than as a conflict of rationalities. The challenge to social psychology, he suggests, is to identify the values that contribute to commons crises. Among those Edney mentions as worth investigating, because they have already been implicated by one or another theory, are the need for identity, stimulation, rank, and competition, survival and longevity, equity, freedom of choice, and social power in the system.

To some, the protection of the commons has the force of a moral imperative. The idea that one should do as one would have others do is a very old one, common to many religions, and still a forceful idea, even if honored more in word than in action. From this perspective, free-loading on commons assets can be viewed as a form of cheating, and morally reprehensible behavior (Brubaker, 1975). From this perspective it is easy too to see behavior that is protective of common assets as rational even given a maximization-of-expected-utility standard of rationality, because personal respectability can be seen as an important component of the utility that one is trying to maximize.

Another way of dealing with social dilemmas, or the prisoner's dilemma more generally, which is hinted at in some of the foregoing discussion, is that of cultivating the ability to take pleasure in advancing the well-being of people other than oneself. I have in mind something more than generosity as enlightened self-interest—the idea that one should be generous because one's behavior toward others tends to evoke similar behavior toward oneself—although that idea has much to commend it. I want to argue the desirability of being able to get satisfaction from other people's successes per se. This is not the most natural of tendencies, especially when the other people involved may be seen as one's competitors in one or another way.

On the other hand, there is ample evidence that we have the capability to empathize and to get immense satisfaction and enjoyment from seeing others benefit, at least in certain situations, even when their benefit has no direct benefit to ourselves. People do engage in deeds of altruism, often at considerable cost to themselves. Some even choose dominated, but corporately beneficial, strategies in social dilemma situations (Hofstadter, 1983). The question is how to increase people's ability or tendency to derive genuine pleasure from contributing to the well-being of others. Of course, if a person derives pleasure from contributing to the well-being of an "opponent" in a prisoner's dilemma situation, the situation may no longer be a dilemma from that person's point of view, because a matrix that took this pleasure into account might show the cooperative choice to be the dominating one. But perhaps we could live with that.

An Infinitely Frustrating Dilemma

The prototypical ethical dilemma is a situation in which one finds oneself forced to make a choice between alternative courses of action each of which requires the violation of an ethical principle to which one is committed. This type of situation is often illustrated by the case of the military leader forced to choose between duty to family and duty to those he commands, under circumstances in which electing to fulfill one duty requires neglecting the other: Both duties are morally binding and neglecting either of them constitutes moral failure, but the choice is forced. Moral dilemmas can also arise when only a single principle is involved, as when, through no fault of one's own, one finds it necessary to choose between which of two promises to keep when keeping one means necessarily reneging on the other (Marcus, 1980).

Slote (1986/1990) has raised the question of whether rational dilemmas that are analogues to these types of moral dilemmas are possible. He confesses to being unable to think of one that fits the first model but gives the following imaginary situation as an example of the second:

> Imagine a science-fictionalized fountain of youth with some very special properties. This fountain emits life-and-happiness-giving rays and can work for a given person only once and at a certain precise moment. Depending on how far from the fountain one is at the exact time when its rays bombard one, one will be given additional days of life and happiness. Assume further that one is capable of standing as close as one pleases to the fountain. For any *n*, one is capable of standing 1/*n*th of an inch away from the fountain, and if one stands 1/*n*th of an inch away one will receive *n* extra days of happiness (if one touches the fountain all bets are off). (p. 470)

We should assume also that the individual who has to decide how close to the fountain to stand wants to maximize the number of happy days of life. What makes this a dilemma is the fact that no matter how close to the fountain one chooses to stand one could have chosen to stand closer: "If it is irrational to do one thing, when one has more reason to perform some alternative, when it would have been better for one to have performed some definitely available alternative, then one has, inevitably, acted irrationally in the circumstances just mentioned. We have described a rational dilemma" (p. 470).

Slote (1986/1990) points out that the dilemma he describes is a dilemma only from the perspective that rationality dictates that one attempt to maximize the expected utility of one's decisions, or that one select the best of all possible alternatives when faced with a choice. It is not necessarily a dilemma from a satisficing point of view. For a satisficer it would be rational to take the attitude that it suffices to get close enough to the wall to add many happy days to ones life and that it is not necessary to try to add the maximum possible number. One might even take the position that it is irrational to try to maximize when maximization is impossible, not only practically but theoretically as well.

Here is a variation on the situation described by Slote. Suppose you are competing with several other people for some very highly desirable prize that is to be given to the person who names the largest number. Each of you has only one chance to name a number, and the winner takes all; there are no consolation prizes. It seems reasonable to assume that the larger the number you name, the greater is the likelihood that you will win the prize. But no matter what number you name, one can always ask why you did not name a larger one.

It is not so clear that the satisficer's perspective works as well in this case as in the preceding one; one cannot argue that a given large number will yield a "good-enough" prize, because if it is not the largest number among those named, it will yield no prize at all. The argument that it is not rational to try to maximize when maximization is not possible also does not apply, because one need not maximize anything in this case, one need only pick a number that is larger than that picked by any of the other contestants. Perhaps this is not a dilemma from the perspective of a model of rationality that assumes one should maximize expected utility but that, like Good's (1983a) Type II rationality, discounts utility by taking the labor and cost of calculations and thinking (or fretting) into account.

Slote (1986/1990) considers this interpretation of his version of the dilemma: "The difficulty of standing at closer and closer distances thus approaches the limits of the agent's power and skill. In that case, an agent who through great effort and concentration chose to stand at a distance so near that it would have been very difficult (for him or for anyone else) to choose to stand any nearer, may count as having chosen rationally, or at least not irrationally" (p. 477). A similar argument may be applied to the task of naming a large number. Slote notes that this type of argument would not apply if it could be assumed that standing nearer and nearer the fountain (or naming a larger and larger number) need not get increasingly difficult, but this seems implausible.

Problems like those represented by the prisoner's and other social dilemmas continue to be debated by decision theorists and other social scientists and are not likely to be resolved to everyone's satisfaction any time soon. As Moser (1990) has pointed out, the debates involve important questions about the roles of cooperation and causation in rational decision making. They also raise fundamental issues about human nature, not only regarding the question of what it is but that of what we would like it to be.

Summary

Dilemmas are difficult decision problems, necessitating, as they often do, choices between equally unsatisfactory (or in some cases equally attractive) options. Some—social dilemmas—pit self-interest against the common good. What constitutes rational behavior in dealing with dilemmas has been a question of interest to many theorists and students of human reasoning. Certain prototypical situations—notably various versions of the prisoner's dilemma—have been intensively studied with the hope of gaining insights into the determinants of competitive and cooperative behavior.

What one sees as rational behavior in many of the social situations that have been studied is likely to depend on whether one conceives of rationality as enlightened self-interest in a fairly narrow sense, or one factors in the roles of such variables as self-image and conscience, or one allows for the possibility that a rational person might take a genuine interest in the welfare of other people, independently of his or her own. In other words, what is perceived as a dilemma from one perspective may not be a dilemma as perceived from another; a decision problem that is very difficult when assessed relative to one set of values may be very easy when assessed relative to another set.

This is not to suggest that all dilemmas are easily resolved simply by getting one's perspective right or that, in particular, social dilemmas are invariably resolved simply by developing a social conscience. Even the most magnanimously altruistic among us can be faced with choices that are very difficult because of the options and uncertainties involved. The study of how people deal with contrived dilemmas has yielded interesting and useful information about competitive and cooperative behavior, but much remains to be learned about how people deal with the dilemmas they encounter in life and how their ability to resolve them effectively might be enhanced.

CHAPTER 7

Statistics

Statistics, more than most other areas of mathematics, is just formalized common sense, quantified straight thinking.

—*Paulos (1992, p. 58).*

Few, if any, characteristics of the world are more apparent than variability. We see it everywhere we look. There are, by some counts, tens of millions species of living creatures in the world. When our focus is narrowed to a single species, say our own, variability is still the rule. Excluding identical twins, no two people look exactly alike and even twins can usually be distinguished by people who know them well.

One of the ways in which we cope with diversity is through the process of conceptual categorization. We group things that are similar in certain respects into conceptual classes or categories, give these categories names and then, for many purposes, respond to members of the same category—items with the same name—as though they were identical. Even within categories, however, there is much variability. Though all chairs are chairs, they differ greatly in size, shape, color, and numerous other respects. All snow flakes are six-sided, but no two of them, we are told, have precisely the same crystalline structure. Whereas people share certain characteristics that define their humanity, they differ in height, weight, age, intelligence, hair color, eye color, and countless other less obvious respects.

Statistics is the area of mathematics that helps us deal with variability in a quantitative way. Compared to some areas of mathematics, such as geometry, algebra, trigonometry and even analysis, statistics emerged as a distinct discipline relatively recently. Since its development, however, its influence on thinking in the social, biological, and physical sciences, as well as in the business world and even in everyday life has been profound.

POLITICAL ARITHMETIC AND THE BIRTH OF STATISTICS

Early in the 19th century there was an explosion of interest in counting among Western governments. Births, deaths, marriages, illnesses, church attendance, suicides, and crimes of various sorts began to be tallied with great enthusiasm. This is not to suggest that interest in such numbers was nonexistent before this time. Sometime in the second millennium B.C. Moses had the Israelites counted by tribe and found the number of males 20 years of age and older and fit for military service to be 603,550 (Numbers 1:46). Caesar Augustus had a census taken throughout the Roman empire at about the time of Jesus' birth (Luke 2:1). A table of life expectancies is known to have been developed by a Roman jurist named Ulpian around 225 A.D., and statistical sampling in rudimentary form goes back at least to the 13th century (P. L. Bernstein, 1996).

Elementary demographic data—births and deaths of a populace—were compiled in England by a merchant, John Graunt, in the 17th century and published, with analyses and interpretative commentary (e.g., regarding causes of death), in his *Natural and Political Observations Made Upon the Bills of Mortality.* Others who built upon the work of Graunt included Edmund Halley's application of demographic data to the valuation of annuities. Although annuities had been sold for a long time, the basis for their valuation was less than ideal; Bernstein says that Ulpian's tables were the last word on the subject for more than 1,400 years. He notes that the policy in England was to sell annuities to everyone at the same price and that this continued until late in the 19th century, despite Halley's work on the subject: "After publication of Halley's life tables in *Transactions* [a newly established journal of the Royal Society] in 1693, a century would pass before governments and insurance companies would take probability-based life expectancies into account" (p. 87). The full title of Halley's work was *An Estimate of the Degrees of Mortality of Mankind, Drawn From Curious Tables of Births and Funerals at the City of Breslaw: With an Attempt to Ascertain the Price of Annuities Upon Lives.* Todhunter (1865/2001), among others, credits this document with laying the foundation of a correct study of the value of life annuities.

One of the motivating forces behind the burst of counting and tabulating activities in the 19th century was a growing awareness of the need to address social problems associated with the rapid growth of cities at the time. It was in large measure from this interest in counting for social or political purposes that statistics as a mathematical discipline eventually emerged: "Statistical laws that look like brute, irreducible facts were first found in human affairs, but they could be noticed only after social phenomena had been enumerated, tabulated and made public. That role was well served by the avalanche of printed numbers at the start of the nineteenth century" (Hacking, 1990, p. 3).

By way of justifying his use of the term *avalanche,* Hacking points out that in Germany alone, there were 410 statistical periodicals being published by 1860, whereas as of 1800 there had been essentially none. The tabulations produced by the counters during this period provided practically unlimited opportunities for subsequent workers to search for, and theorize about, statistical regularities in real-world events. The tabulators were to subsequent theorists what Tycho Brahe was to Kepler.

The information and the "political arithmetic" ("social mathematics," "moral statistics") that developed around these numbers reflected a desire to provide a scientific basis for public policy and were used for a variety of governmental purposes, especially to justify political or social reforms. The active members of the early statistical societies were more likely to be politically active than to be seriously interested in mathematics or natural science (Porter, 1986). Although eminent mathematicians were involved as well. Laplace (1814/1951), for example, applied his calculus of probabilities to the "moral sciences" and a variety of social issues; noting that average life expectancy in France rose from about 28 at birth to about 43 for those who had already lived beyond infancy, he argued the case for vaccination as one means of attacking infant mortality. Gigerenzer et al. (1989) describe the development of statistical thinking in the 19th century as reflecting a mix of social and political views: "Even in 1900, after the successful application of statistical reasoning to physics and biology, and the beginnings of a mathematical field of statistics, the term still referred first of all to social numbers, and only by analogy to this branch of applied mathematics" (p. 69).

Interest in tabulating and applying the results to the resolution of social problems appears to have become somewhat weaker during the second half of the 19th century than it had been during the first half. And little progress was made in developing social statistics into a formal discipline during this time: "Fifty years after the start of the era of statistical enthusiasm in the 1820s, the same impressionistic and arbitrary eyeballing techniques were used to argue for a positive or negative relationship between two variables. No permanent cumulation in techniques of data collection and analysis had been made. Until the twentieth century, this is typical of empirical social research everywhere" (Oberschall, 1987, p. 113). "By the end of the [nineteenth] century," Oberschall says, "moral statistics meant voluminous compilations of statistical data without an attempt to make sense of its contents in other than a superficial descriptive fashion" (p. 115).

But statistics did become a mathematical discipline in time, and its effect on thinking extends far beyond political arithmetic. The work of Quetelet and other early tabulators or "statists" called attention to the possibility of studying and describing aggregate phenomena, and even of developing laws in terms of which such phenomena might be understood, despite the unpredictability or "unlawfulness" of the individual cases of which the aggregates were composed. The ideas and methods that were developed were found to be useful in diverse domains in addition to government and public policy, including especially insurance and biometrics, and eventually the physical sciences as well. Statistics is one area of mathematics in which applications often spurred theoretical developments. Stewart (1989) points out that by the end of the 19th century statistics had provided an alternative to differential equations as a basis for describing many natural phenomena, but that despite this fact there was virtually no contact, at a mathematical level, between the two disciplines.

During the first half of the 20th century, mathematical statistics developed rapidly and affected science profoundly. Scientists had been conducting experiments long before this time, but before the 20th century the results of experimentation were seldom reported very fully; instead, scientists typically presented their conclusions and published data that "demonstrated" their truth. The work of a few statisticians—notably Ronald Fisher—produced an approach to experimental design and data analysis that changed the way experimental science was done (Salsburg, 2001).

Fisher's work had a profound effect on psychological research, especially on the design of experiments and the interpretation of results. Statistical hypothesis testing became the standard modus operandi. But during the latter half of the 20th century, statistics began to affect the thinking of psychologists in another important way. As Gigerenzer et al. (1989) put it

> The view that the nature of human thought might be statistical calculation, or at least, should be, arose at the same time as the view that sensory detection and discrimination, perception, recognition from memory, and other cognitive processes might involve statistical calculation.... All these statistical perspectives emerged around 1960, shortly after statistics had been institutionalized as the indispensable tool of the experimenter. (p. 216)

In physics, probabilistic and statistical ideas became important in two different ways; first, during the latter part of the 18th and the beginning of the 19th century, in application to the treatment of errors of observation, and later, during the latter part of the 19th century and early in the 20th, in theory construction (Krüger, 1987a). These ideas impacted theory construction first in statistical mechanics and then in quantum mechanics. The application of statistics to biology was spurred by the rediscovery of Mendel's work on heredity in 1900 (R. J. Berry, 1988). This application met with some resistance because not only did Mendel's theory run counter to prevailing ideas about inheritance, but the application of mathematics to biology was itself unacceptable to many biologists of the day (Barber, 1961; Flammarion, 1890). In time, however, statistical analysis came to be seen as an indispensable tool for biologists, as well as for science more generally. By the end of the 20th century, probability and statistical ideas had permeated essentially all the sciences, hard and soft alike. Perhaps nowhere is the usefulness of statistics in science more apparent than in the intersection of astro- and particle physics, where the tenability of theories of cosmology rest, in part, on the ability to estimate the frequency of very low probability events (e.g., the splitting of a deuteron nucleus by a neutrino) by applying statistical techniques to very large numbers of observations (MacDonald, Klein, & Wark, 2003).

STATISTICAL REGULARITY OR THE LAW OF LARGE NUMBERS

> When dealt with in sufficient numbers matters of chance become matters of certainty. (Shaw, 1944/1956, p. 1525)
>
> In some respects a crowd of phenomena is more easy to manage than a few individuals. For a certain order is generated by chaos. (Edgeworth, 1890, p. 471)
>
> It is almost impossible to study any type of life without being impressed by the small importance of the individual. (Pearson & Weldon, 1901, p. 3)

As we have seen, random does not mean unlawful, or even unpredictable. Though one cannot reliably predict the outcome of a single toss of a fair coin or a single roll of a fair die, one can confidently predict that the percentage of heads in a large number of tosses will be close to .5 and the percentage of sixes in a large number of throws of a fair die will be close to .167. The "transition from uncertainty to near certainty" as we change our focus from individual events to large aggregates of events is, as Ruelle (1991) puts it, "an essential theme in the study of chance" (p. 5). It is because of such aggregate regularity and predictability that we can talk about the "laws of chance."

Polya (1954b) recognizes this fact in his reference to the theory of probability as "the theory of certain observable phenomena, the *random mass phenomena*" (p. 55), that are characterized by aggregate regularity despite the unpredictability of individual happenings. Rainfall, for example, is a random mass phenomenon: One can make reasonable predictions about how much rain will fall in a specified area during a specified time, given a shower of specified intensity, but one cannot predict with any certainty where the next drop will fall.

The idea that regularities may be seen in the behavior of aggregations that are not apparent in the behavior of the elements of which the aggregations are composed grew out of observations of such regularities in the world. One of the first such regularities to be noticed was the fact, discussed by John Arbuthnot in 1710, that the ratio of male births to female births was consistently greater than 1 by a very small amount. No one could predict with any consistency the gender of children to be born on a case-by-case basis, but one could be quite sure that if one checked a large number of births over a period of time the number of boys born would exceed the number of girls by a small percentage.

When people examined the statistical data that governments began gathering in great quantities in the early 19th century, they noticed a number of other constancies, some quite surprising. Quetelet (1829), for example, reported being shocked at the "frightening regularity" with which the same crimes were committed year after year. Births, marriages, and deaths by age were also observed to be astonishingly regular, and hence predictable, from year to year. Studies of suicides revealed not only remarkable constancy in the overall rate from year to year, but constancy in the relative frequency with which different methods were used within a given location (a specific method was used with different relative frequency in London and Paris, for example, but with close to the same relative frequency from year to year in each place). Laplace (1814/1951) even reported that the number of dead letters in the Paris postal system was relatively constant from one year to the next.

Although statistical regularities were first noticed primarily in social and behavioral data, once the idea of their existence had emerged, investigators began to look for them everywhere. And it appears that they found them wherever they looked. The 19th-century work that was done on statistical regularity was empirical in the extreme: "In the human and social arena, and more generally in the whole domain of the nascent concept of statistical law, it was the Baconian generalizers who did the work. They were ready and willing to produce 'laws' when they had no more theoretical understanding than Quetelet had of Belgian lilacs" (Hacking, 1990, p. 62). (Quetelet discovered the law of blooming lilacs, according to which lilacs in Brussels bloom when the sum of the squares of the mean daily temperatures since the last frost totals 4264°C^2.)

In addition to providing grist for the mills of future theorists, the enormous quantities of data that were collected during the 19th century served some useful functions. Regarding the effect of these data on public health, Hacking (1987b) has this to say: "The marvels of modern medicine produce modest increases in life expectancy that are peanuts compared to that coconut of an increase provided by the sanitary movement and its band of well-meaning statisticians" (p. 51). Also noting its implications for public health and elementary education, Metz (1987) calls the statistics of the time "a social science for practical men" (p. 340): "The formation of sanitary statistics as the major event of the statistical movement and the development of a policy of public health demonstrates the important role that ideological factors played in the transformation of statistics into a strategy of social amelioration" (p. 344).

An interesting statistical regularity is the inverse power law, according to which some function of a variable varies inversely with a power of that variable. An example of this type of relationship is the inverse square law proposed by Alfred Lotka (1926) as a description of the frequency distribution of scientific productivity; he found that the number of scientists who produced *n* papers was inversely proportional to n^2. Price (1961) has suggested that an inverse-square-law also describes the relationship between the quality of scientific papers (as indicated by the number of times they are cited by other scientific publications) and the number of such papers produced, which is to say that for every paper of first quality, there are four of second rate, nine of third, and so on.

Another example of the same type of law was proposed by Pareto (1897) to describe the function relating number of people having an income of size *n* and *n;* in this case, using $I(n)$ to represent the number of people with income of size n, the relationship is approximately $I(n) = 1/n^{1.5}$. In a study of uses of a corporate electronic bulletin board, I found that the number of people who posted *n* messages fell off roughly with $T/2^{n-1}$, where *T* is the number of people who posted at least one; this means that about half as many people posted two messages as posted one, about half as many posted three as posted two, and so on (Nickerson, 1994). This is an example of the Poisson distribution where each value of the variable is a constant multiple of the preceding value; it illustrates the occurrence of statistical regularity in what might appear to be random behavior. The Poisson distribution has been found to be descriptive of behavior in many contexts. When the number of instances in which *n* different scientists have made the same discovery more or less simultaneously is plotted against *n,* for example, a Poisson distribution is obtained (Merton, 1961). Distributions of these types are reminiscent of the general inverse relationship found by Zipf (1949) between the number of occurrences of something in the world and the position of that thing in a list ordered by size.

The observed stability of statistical aggregates provided the foundation for the "social physics," with its central concept of *l'homme moyen,* or the average man, that Quetelet established in 1831. It also provided an empirical basis for the "law of large numbers," which was articulated by Poisson (1837) in reference originally "to the tendency for events frequently repeated and not too closely dependent on one another—that is to say, virtually everything counted by governmental statistical agencies—to occur in approximately constant numbers from year to year" (Gigerenzer et al., 1989, p. 40). The work of Quetelet (1847, 1848), Buckle (1857–1861), and others did much both to call attention to the phenomenon of statistical regularity in human affairs and to stimulate debate on its philosophical significance. Quetelet's critics believed he sometimes saw regularities that did not exist (Porter, 1987), but his work was exceptionally influential nevertheless. Oberschall (1987) describes his writings this way: "If one skips the pages on celestial mechanics, the average man, social physics, and the center of gravity, Quetelet's writings are filled with a wealth of observations, comments, suggestions, and facts about crime, marriage, suicide, and dozens of other topics, which are quite instructive. Many must have enjoyed reading him for this reason and not because of his philosophic and methodological pronouncements" (p. 112).

The law of large numbers, as the term is currently used by statisticians, has a somewhat narrower connotation. It says that the larger the random sample, the more closely the relative frequency of chance events will conform to the mathematical probability of those events. Stated slightly differently, it says that one can make the difference between the theoretical probability of an event and the relative frequency of that event arbitrarily small simply by making the sample sufficiently large. Thus, given a fair die, for which the probability of rolling a three is 1/6, the larger the number of times the die is rolled, the smaller the difference will be between the relative frequency of the occurrence of a three and 1/6. It is to be noted that the law does not say that the difference between the actual and predicted *numbers* of events will decrease—this difference is likely to increase with sample size—it says only that the difference between the theoretical probability and the obtained ratio will decrease.

It is not necessary to know why the law of large numbers works, or even to consider the question, in order to use it to good effect. One may take it as a general principle that has been found to be descriptive of the behavior of variables in a large number of instances, and let it go at that. But of course, such regularity begs an explanation. Why should regularity result from the combining of many irregular events? Why should things be predictable in the aggregate when they are completely unpredictable individually? Does the predictability of human behavior in the aggregate have implications for beliefs about determinism, free will, and moral responsibility?

The fact that the practical usefulness of statistics does not depend on having answers to such questions does not make them uninteresting, and such questions have intrigued thinkers since the phenomenon of the regularity of aggregations of irregular events was first noticed. Some have seen in statistical regularity what Hacking (1987a) calls a "grim determinism" (p. 382). The idea, promoted especially by Buckle (1857–1861) was anathema to social reformers because it implied the futility of efforts at social reform, and reform was a main interest of many of those who gathered or paid the most attention to the numbers. The idea of statistical determinism was rebutted hard and often and appears to have had little effect on efforts at social change.

The predictability—the aggregate regularity—of phenomena that result from the combined behavior of very large numbers of component entities still begs an explanation, whether the components are assumed to be predictable individually or not: "Events and actions such as births, deaths, marriage, and crime manifest regular patterns in the aggregate. Yet these patterns result from the uncoordinated activities and choices of a multitude of people, each pursuing a private end, and not from an imposing design. How is such a spontaneously generated orderliness at all possible?" (Oberschall, 1987, p. 104). The question has been asked countless times, and it remains as valid a question today as when it was first asked.

Consider again that prototypical example of a chance process, a coin-tossing experiment. Although the situation is sufficiently complex that prediction of the outcome of any given toss is very difficult, we assume the outcome of the toss of a fair coin is determined by the laws of physics. Why, in a very large number of tosses, should the forces that determine the outcome of every toss balance themselves out so completely that each of the two possible outcomes occurs almost exactly the same proportion of times? To say that the individual causal factors are equally likely to change in one way as in another from toss to toss is little help, because the question then becomes, why should this be so? And if the individual event were not predictable in principle—if the outcome of the individual toss were assumed to be a true fortuitous event, in Poincaré's sense—we would still have the question of why a large number of such events is so reliably well-behaved in the aggregate. The lawfulness of chance is an enigma.

Several 17th- and 18th-century writers, including Arbuthnot (1710) and De Moivre (1756/1962), saw aggregate regularities as evidence of design and the work of a Designer. Here is De Moivre's conclusion on the matter:

> As it is thus demonstrable that there are, in the constitution of things, certain Laws according to which Events happen, it is no less evident from Observation, that those Laws serve to wise, useful and beneficent purposes; to preserve the steadfast Order of the Universe, to propagate the several Species of Beings, and furnish to the sentient Kind such degrees of happiness as are suited to their State.
>
> But such Laws, as well as the original Design and Purpose of their Establishment, must all be *from without:* the *Inertia* of matter, and the nature of all created Beings, rendering it impossible that any thing should modify its own essence, or give to itself, or to any thing else, an original determination or propensity. And hence, if we blind not ourselves with metaphysical dust, we shall be led, by a short and obvious way, to the acknowledgment of the great MAKER and GOVERNOUR of all; *Himself all-wise, all-powerful* and *good.* (p. 264)

Regularities revealed through statistical research continued to evoke expressions of amazement by 18th- and 19th-century writers. Pointers to several examples of them are given by Porter (1986). Even Galton, who would not have agreed with De Moivre's conclusion, was greatly impressed with "the wonderful form of cosmic order expressed by the 'Law of Frequency of Error,'" which he referred to as "the supreme law of unreason." He wondered, for example, about why it is that successive generations of people are so alike in the aggregate despite the fact that individual offspring typically differ considerably from their parents: "Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. The tops of the marshaled row form a flowing curve of invariable proportions; and each element, as it is sorted into place, finds, as it were, a preordained niche, accurately adapted to fit it" (from *Natural Inheritance,* quoted in Porter, 1986, p. 146). Such admissions of wonderment are less in evidence among 20th-century writers. Is it because we have a more sophisticated attitude on the matter? Has familiarity with the phenomena bred contempt? Could it be that our lack of amazement is sometimes the consequence simply of a lack of thought?

The idea that statistical regularity is the result of design, or is maintained by the Designer, was attacked by Nicholas Bernoulli, d'Alembert, and Poisson, among others (Daston, 1987b), and it is not popular among modern theorists, but it has not been replaced with a demonstrably more adequate explanation of the regularity that is everywhere observed. Bernoulli claimed to have refuted the argument as put forth by Arbuthnot, but his refutation consists in assuming that the probability that a random birth will be a boy is 18/35 and showing that his uncle Jacob's limit theorem (see chap. 1) implies that of a large number of children born, the probability that the number of boys will lie between specified bounds can be stated with considerable precision. It does not account for why the probability of a random birth being male should be 18/35. Often in discussions of statistical regularity *chance* is spoken of as a cause, but it is not at all clear what it could mean for chance to be the cause of anything. In saying that something is caused by chance, one is really only saying that something has certain properties, such as statistical regularity.

There is something paradoxical, almost oxymoronic, about the idea of "laws of chance." Chance events are, by definition, indeterministic, erratic, unpredictable. It is a fact that events that are unpredictable individually can be highly regular and predictable in the aggregate and it is this fact that underlies reference to the laws of chance. But the aggregate regularity itself begs an explanation. It is easy to delude ourselves into thinking that, by attributing statistical regularity to the laws of chance we have explained it, when in fact we have simply called it by another name.

VARIABILITY AND "ERROR"

The observation of statistical regularities was critical to the development of statistical thinking; so also, however, was the observation of variability. For many purposes the variability that was observed, as for example in successive measurements of the same thing, was considered a nuisance—something to be minimized by careful measurement and compensated for after the fact by appropriate analyses. A hint of the idea that variability represents an observational distortion of a Platonic reality is seen in the use of such terms as *error* and *deviation* in its description.

Among the most important constructs relating to variability, from a historical perspective, was the "error curve," or what is referred to more commonly today as the Gaussian or normal distribution or density function. Mathematically, it is the curve defined by

$$f(x)=\left(\frac{n}{\sigma\sqrt{2\pi}}\right)e^{\frac{-(x-\mu)^2}{2\sigma^2}},$$

where n is the number of measurements and μ and σ^2 are their mean and variance respectively. It is descriptive, or approximately so, of a great many distributions found in nature, a fact that Fourier (1819), among others, found to be remarkable. It was applied to probability theory by de Moivre as the limit of the binomial distribution,

$$b(n)=\binom{n}{k}p^k(1-p)^{n-k}.$$

The fact that it can be used to approximate the binomial distribution for large n facilitates the estimation of probabilities that would be computationally tedious if not intractable with the binomial function and is a great convenience in the practice of statistical hypothesis testing and decision making more generally.

Porter (1986) points out that the history of this curve is practically coextensive with the history of statistical mathematics during the 19th century. He also notes that although the curve originally was thought of as a representation of error—it was used effectively by astronomers, along with the method of least squares, to deal with the problem of variability in the measurement of astronomical phenomena—it was reinterpreted as a law of genuine variation. Porter refers to this reinterpretation as "the central achievement of nineteenth-century statistical thought" (p. 91).

The reinterpretation came about gradually by a process that in retrospect looks like creative misunderstanding:

> Once again the lead was taken by social thought, and in this matter its influence on the natural sciences is demonstrable. The main line of development proceeded from Laplace to Quetelet to Maxwell and Galton, and from the error of mean values in demography as well as astronomy to the deviations from an idealized average man to the distribution of molecular velocities in a gas and the inheritance of biological variation in a family. Ultimately, even the analysis of error was transformed by this line of statistical thinking. (Porter, 1986, p. 91)

Thus over time, beginning perhaps with Quetelet's discovery that the "error curve" was descriptive not only of measurements made by astronomers but also of such natural variables as human height and other features, there developed an interest in variability as a natural phenomenon in its own right. Around 1840, Quetelet's focus began to shift from the stability of averages and rates to the way variables are distributed. He was particularly intrigued by how often the distribution appeared to be similar to that of errors of observation (Gaussian) and in his 1848 book, *Du Système Social et des Lois Que le Régissent (On the Social System and the Laws That Govern it)*, proposed the "law of accidental causes," to account for the regularity of distributions of all sorts (Lécuyer, 1987).

Interest in variability per se, and especially in the variability of human traits and abilities, was apparent in the work of Galton (1869, 1874, 1889) and his lifelong focus on the exceptional as opposed to the commonplace. His study of genius and his framing of the nature–nurture debate reflect this interest, and in his autobiography, he explicitly acknowledges it: "The primary objects of the Gaussian Law of Error were exactly opposed, in one sense, to those to which I applied them. They were to get rid of, or to provide a just allowance for errors. But those errors or deviations were the very things I wanted to preserve and to know about" (Galton, 1909, p. 305). Murray (1987) identifies Fechner's use of the equation for the normal distribution to estimate the variability in measurements in psychophysical experiments on just noticeable differences as the first application of probability theory in psychology.

The history of the development of statistical thinking—extensive accounts of which are readily available (e.g., P. L. Bernstein, 1996; David, 1962; Gigerenzer et al., 1989; Hacking, 1975, 1990; Porter, 1986; Salsburg, 2001)—presents a curious mix of discoveries of lawful regularities in the behavior of large collections of unpredictable individual cases, frustrations arising from the inability in certain situations to get precisely the same measurements on different attempts to measure the same thing, the realization that certain distributions are characteristic of a wide variety of natural phenomena, interest in why things vary as they do, and the practical need for mathematical concepts and tools with which to deal with variability in an effective way. The purpose of statistical hypothesis testing, about which more later, can be seen as that of distinguishing between variability (in one or more dependent variables) that can be attributed to controlled variability (in one or more independent variables) and that which cannot. In this context, error generally connotes the variability in the first kind of variable that cannot be attributed to the variability in the second kind, and it is sometimes divided into two types: sampling error (error arising from studying a sample that is not representative of the population of interest) and measurement error (error that results from imprecision of a measuring instrument) (Abelson, 1995).

USES OF STATISTICS

> After the mid-nineteenth century, it became common to investigate collective phenomena using what came to be called the statistical method, the method of reasoning about events in large numbers without being troubled by the intractability of individuals. (Porter, 1986, p. 12)

Today statistics is a well-established area of mathematics with countless areas of application. In the brief overview of the topic that follows, it will be convenient to distinguish four specific purposes that statistics serves: description, estimation, hypothesis testing, and explanation.

Statistical Description

Two of the more fundamental ideas in statistics are those of a variable and a frequency distribution. A variable is any measurable property that varies in a population. Height, for example, is a variable property of humans. Variables are said to be distributed in certain ways. If we were to measure the height of every person in the world and then construct a graph showing the number of people whose height was between, say, 5 ft. and 5 ft. 1 in., those between 5 ft. 1 in. and 5 ft. 2 in., those between 5 ft. 2 in. and 5 ft. 3 in., and so on, when we had finished, extending the height measure in both directions sufficiently far to accommodate the shortest and tallest people in the world, the resulting graph would represent the distribution of the height of all people living in the world today. This distribution would be more or less bell-shaped with its highest point probably somewhere around 5½ feet.

Not everything is distributed like height of course. If we were to plot the distribution of family sizes, for example, showing the number of children in the family, varying from 1 to, say, 20, on the horizontal axis and number of families with that number of children on the vertical axis, we would find that the distribution would be quite asymmetrical. There would be many more families with 1 or 2 children than with 19 or 20.

If we were to plot the distribution of many variables, we would find that different variables give rise to distributions with different shapes. However, we would also discover that certain shapes occur frequently, because there are a few distributions that are descriptive of the ways in which many variables are distributed in the world. Among the better known of these, in addition to the gaussian, are the rectangular, the exponential, and the logistic.

Producing a picture of a distribution is an excellent way to answer a question regarding how any particular variable is distributed. However, although pictures are very useful for some purposes, they are less so for others. Statisticians have developed concepts that can be used to describe distributions when one wants to talk about them. These descriptions make use of properties or characteristics of distributions called "parameters." A parameter of a distribution is a number that provides some summary information about the distribution. The mean, median, and mode are sometimes referred to as central-tendency parameters, because each of them conveys information about the center, more or less, of the distribution. Other useful parameters include the standard deviation and interquartile range, which, because they convey information about the shape of the distribution and, in particular, about the degree to which it is narrow and peaked as opposed to broad and relatively flat, are sometimes referred to as parameters of dispersion.

Measures of central tendency and of dispersion are very useful in describing distributions. Thus, in answer to the question of how tall American men are, one might point to a graph of the distribution of heights or, alternatively one might say that this distribution is approximately gaussian, or "normal," and has a mean of 68 inches and a standard deviation of 3 inches. To someone who understands elementary statistics, such a description conveys nearly as much information as does the actual distribution graph. Such a person understands that a gaussian distribution has the approximate shape of a bell, symmetrical about its mean, and that about 95% of all the cases are within plus or minus two standard deviations of the mean. In the case of our example, the description tells us that the average height of an American man is 68 inches and that about 95% of all American men have a height somewhere between 62 inches and 74 inches. (The numbers in this example were made up for purposes of this illustration.)

A particularly useful fact relating to the normal distribution is represented by what is known as the *central limit theorem*. According to this theorem, no matter how a variable is distributed, if one draws a large number of random samples of that variable and calculates the mean of each sample, one will find that the distribution of the means will be normal. "Large number" in this context is conventionally taken to be 30 or more, unless the underlying distribution is highly asymmetric, in which case it may have to be a bit larger for the approximation to be accurate. This is a remarkable theorem and one that finds many practical applications in statistics.

Distribution parameters facilitate discussion about variables, but it is important to recognize that they convey less information about variables than do the distributions themselves, in either graphical or equation form. Two distributions with the same mean can differ with respect to dispersion; and even if they have equal standard deviations, the variability may be determined in different ways.

The usefulness of statistics for purposes of description does not depend on relating statistics to any theory of probability or chance or on any assumptions about why natural variables are distributed the way they are. But one would like to know why so many variables are distributed in ways that can be described by simple mathematical functions. What determines how a variable is distributed? Why are certain distributions so frequently seen in nature? What accounts for instances in which two variables that appear to have no relationship to each other are distributed in the same way?

Statistical Estimation

Often it is not practical or perhaps even possible to determine how a variable is distributed in some population of interest. In that case, what we can do is determine how the variable is distributed in a subset of the population that is assumed to be representative of the population as a whole. Such a representative subset is referred to as a sample and measurements made on it are taken as estimates of what those measurements would be if made on the entire population from which the sample was drawn.

The use of statistics for purposes of estimation is a highly developed methodology. There are well-tested procedures for ensuring that samples are indeed representative of the populations from which they are drawn. Estimates of population parameters typically are accompanied by statements of the margin of error of those estimates; these statements (which are themselves estimates) often indicate a range of values within which the true value is believed to be highly likely to lie. In general, the precision of an estimate (narrowness of margin of error) and the confidence expressed in it vary directly with the size of the sample on which the estimate is based and inversely with the variability of the values in it.

Many socially significant decisions are made on the basis of actuarial statistics. Life insurance premiums are determined by life expectancy tables that show how much longer people who have attained a specific age are expected to live, on the average. Academic and job placement decisions are made, in part, on the basis of the relationship between performance on psychological tests and academic or on-the-job performance of large numbers of individuals who have taken those tests and performed those jobs in the past. Use of statistical reasoning is justified economically on the grounds that the resulting decisions are better, in the aggregate, than they would be if such reasoning were not used. Calibrating life insurance premiums to average life expectancy means that the highest premium costs are borne by people whose policies are likely to be cashed in within the shortest time. Careful use of standardized tests for placement purposes presumably increases the relative frequency of appropriate placements.

What is true of a population need not be true of any single member of it; thus whereas, as of 1990, the average life expectancy of an American female at birth was 78.8 years, relatively few American females who were born in 1990 will live precisely 78.8 years. Though it is true that the incidence of automobile accidents is much higher for male teenage drivers than for all other drivers, it is not the case that every male teenage driver is accident prone. Decisions regarding individuals that are based on statistical data that are descriptive of the population to which those individuals belong are often unfair to the individuals. Does that make them irrational? The answer must depend, of course, on how one defines rationality and applies the definition to this particular case.

A special instance of statistical estimation that can have important social consequences is that of opinion polling. On the basis of the sampling of relatively small numbers of people (often a few hundred) pollsters present estimates of how the general populace, or some specified subset of it (e.g., members of an ethnic group, a political party, an age category) feel, or what it believes, about particular issues. There are many problems associated with opinion polling that have little to do with statistics, not the least of which is the sensitivity of the outcomes to the specific wording of the questions asked and the consequent vulnerability of polls to manipulation, but there are many statistical concerns as well. The problem of representative sampling is a crucial one, as is that of the possibility of samples being biased as a consequence of participant self-selection. Many polls relate to matters about which not all people are equally willing to express their personal feelings or beliefs; especially when this is the case, there may be good reason to doubt that the opinions expressed by people who are willing to participate are representative of those of people who are not (Moore, 1992).

Statistical Hypothesis Testing

> The mathematical statistician has become a universal expert, whose specialty is not so much a subject matter as a method of inference applicable to all subject matters. (Gigerenzer et al., 1989, p. 69)

> Random sampling and random designs of experiments were introduced into statistics to achieve apparent precision and objectivity. I believe that the precision attained by objectivistic methods in statistics invariably involves throwing away of information. (Good, 1983a, p. 86)

The term "null hypothesis" has at least two connotations as applied to statistical testing. Usually it is intended to represent the hypothesis of "no difference" between two sets of data with respect to some parameter, such as their means, or of "no effect" of an experimental manipulation on the dependent variable of interest. Suppose one wished to know whether Europeans and Asians differ with respect to how long they live. To approach this question statistically one would compare representative samples of longevity data from Europe and Asia. The specific question one would ask, by means of the application of statistical procedures, would be whether the evidence justifies rejection of the "null" hypothesis, typically represented by H_0, that the two samples of data were drawn from the same population. This may seem like a strange way to describe the situation because it is obvious that the data are from two different populations, one European and one Asian, but the question that is being asked is if those populations differ with respect to the variable of interest, namely longevity; the default—null—hypothesis is that they do not differ in this regard. Statistical testing is used to see if the data warrant rejection of that hypothesis, or whether it is more reasonable to consider them the same population with respect to this variable.

"Null hypothesis" also sometimes has the more inclusive meaning of the hypothesis the nullification of which, by statistical means, would be taken as evidence in support of a specified alternative hypothesis. Given the latter connotation, the null hypothesis may or may not be a hypothesis of no difference or of no effect (Bakan, 1966).

The distinction between these connotations is sometimes made by referring to the former as the nil-null hypothesis or simply the nil hypothesis; usually the distinction is not made explicitly and whether null is to be understood to mean nil-null, as it almost always does, must be inferred from the context. In what follows, "null hypothesis" will be used to indicate the hypothesis of no difference or no effect unless otherwise specified, and will be represented as H_0.

Application of a statistical significance test to the difference between two means usually yields a value of *p*, the theoretical probability that a difference of the size obtained, or larger, would have been obtained from two samples of the size of those used had they been drawn at random from the same population. A "confidence level," usually designated as *alpha*, is specified to serve as a decision criterion and the null hypothesis is rejected only if the value of *p* yielded by the test is not greater than the value of *alpha*. If *alpha* is set at .05, say, and a significance test yields a value of p less than .05, the null hypothesis is rejected and the result is said to be statistically significant at the .05 level.

This logic of null hypothesis significance testing, as it is generally presented in statistics texts, admits of only two possible decision outcomes: rejection (at some specified level of confidence) of the hypothesis of no difference, and failure to reject this hypothesis (at that level). Given the latter outcome, all one is justified in saying is that a statistically significant difference was not found; one does not have a basis for concluding that the null hypothesis is true.

Inasmuch as the null hypothesis may be either true or false and it may either be rejected or fail to be rejected, any given instance of null hypothesis testing admits of four possible outcomes as shown in the following Table 7.1

There are two ways to be right: rejecting the null hypothesis when it is false (when the samples were drawn from different populations) and failing to reject it when it is true (when the samples were drawn from the same population). And there are two ways to be wrong: rejecting the null hypothesis when it is true and failing to reject it when it is false. The first of these two ways to be wrong is usually referred to as a Type I error and the second as a Type II error. Theoretically, the probability that a Type I error will be made, *if the null hypothesis is true,* is given by the p value (criterion or confidence level) that is used to reject the hypothesis (alpha). The probability of occurrence of a Type II error, *if the null hypothesis is false,* usually referred to as *beta,* is generally much larger than alpha, but not known precisely.

Null hypothesis significance testing (NHST) has been widely used in psychological research since its invention and its use has been subject to intense criticism for equally as long. Debate between critics and defenders of such testing continues to the present time. I have reviewed the controversy elsewhere (Nickerson, 2000), so will not do so again here, beyond noting some of the critics and defenders of NHST and listing in Table 7.2 some common misunderstandings about it that have been pointed out in the literature.

TABLE 7.1

The Four Possible Outcomes of a Null Hypothesis Test

	Truth State of H_0	
Decision re H_0	*False*	*True*
Rejected	Correct rejection	Type I error
Not rejected	Type II error	Correct non rejection

TABLE 7.2

Common False Beliefs About Null Hypothesis Significance Testing

- p is the probability that the null hypothesis is true and that 1 – p is the probability that the alternative hypothesis is true.
- Rejection of the null hypothesis establishes the truth of a theory that predicts it to be false.
- A small p is evidence that the results are replicable.
- A small p means a treatment effect of large magnitude.
- Statistical significance means theoretical or practical significance.
- Alpha is the probability that if one has rejected the null hypothesis one has made a Type I error.
- The value at which alpha is set for a given experiment is the probability that a Type I error will be made in interpreting the results of that experiment.
- The value at which alpha is set is the probability of Type I error across a large set of experiments in which alpha is set at that value.
- Beta is the probability that the null hypothesis is false, or the probability of making a Type II error.
- Failing to reject the null hypothesis is equivalent to demonstrating it to be true.
- Failure to reject the null hypothesis is evidence of a failed experiment.

Critics of NHST include Bakan (1966), Brewer (1985), Chronbach (1975), J. Cohen (1994), Dracup (1995), Falk (1986), Falk and Greenbaum (1995), Folger (1989), Gigerenzer and Murray (1987), Grant (1962), Guttman (1977), Kirk (1996), Lunt and Livingstone (1989), Lykken (1968), Meehl (1967), Morrison and Henkel (1970), Oakes (1986), Pedhazur and Schmelkin (1991), Pollard (1993), Rozeboom (1960), Sedlmeier and Gigerenzer (1989), Shaver (1993), Shrout (1997), and B. Thompson (1996, 1997). Some critics have argued that progress in psychology has been impeded by the use of significance testing, as it is conventionally done, or even that such testing should be banned (Carver, 1978, 1993; Hunter, 1997; G. R. Loftus, 1991, 1995, 1996; Schmidt, 1992, 1996). Defenders of NHST include Abelson (1995, 1997a, 1997b), Baril and Cannon (1995), Chow (1987, 1988, 1989, 1991, 1996, 1998a, 1998b), Cortina and Dunlap (1997), Cox (1977), Dixon (1998), Frick (1996), Giere (1972), R. J. Harris (1997), Kalbfleisch and Sprott (1976), Mulaik, Raju and Harshman (1997), D. Robinson and J. Levin (1997), Sohn (1998), Tukey (1991), W. Wilson, H. L. Miller, and Lower (1967), and Winch and Campbell (1969).

Defenders of NHST testing generally acknowledge that it has limitations and that it is subject to misunderstanding and misuse. They tend to believe, however, that the major problem in its use is not inherent to the technique but lies in misapplications of it, or misinterpretations of the results obtained with it. Abelson (1995, 1997a), for example, defends the use of NHST, but contends that it must be used judiciously, not as an unequivocal determinant of what is worth reporting, but as a means of helping to justify claims that specific effects were unlikely to have been obtained by chance. The worthiness of an experimental finding, he argues, is determined by several considerations, including the magnitude, generality, and interestingness of the effect. He contends, especially in Abelson (1995), that all statistics should be treated as aids to principled argument. Tukey (1991) argues that it is really the direction of an effect, rather than the existence of one, that the *t* test helps decide, and that acquired significance levels should be seen as a guide to whether an effect has been demonstrated but not taken as the sole criterion.

The bottom line is that statistical testing cannot be done with complete objectivity without running the risk of obtaining nonsensical results. The importance of human judgment in the interpretation of experimental results—and of the outcomes of statistical tests—has been stressed by many writers (Abelson, 1997a; Berger & D. A. Berry, 1988; Browne & Cudeck, 1992; J. Cohen, 1994; Cortina & Dunlap, 1997; Falk & Greenbaum, 1995; Gigerenzer, 1993; Huberty & Morris, 1988; Malgady, 1998). Statistical tests are tools that must be used with care. Used judiciously, they can be a great help in making sense of data, but they are easily misused, and by themselves they can never determine whether a result is worthy of attention.

STATISTICAL EXPLANATION

> The omnipresent hypothesis of randomness is an alternative to any other kind of explanation. This seems to be deeply rooted in human nature. "Was it intention or accident?" "Is there an assignable cause or merely chance coincidence?" Some question of this kind occurs in almost every debate or deliberation, in trivial gossip and in the law courts, in everyday matters and in science. (Polya, 1954b, p. 95)

> There are laws of chance. We must avoid the philosophically intriguing question as to why chance, which seems to be the antithesis of all order and regularity, can be described at all in terms of laws. (W. Weaver, 1950, p. 44)

If one plots the frequency with which there have been zero, one, two, three, four, or more than four outbreaks of war in a year, one gets a Poisson distribution. One gets the same type of distribution if one plots the frequency of years having zero, one, two, three, four, or more than four wars coming to an end. (The fact that one gets a similar distribution in both cases would be redundant if all wars lasted the same amount of time, but they do not.) This is the kind of distribution one would expect if the initiation of war were a completely random event. It suggests, in effect, that the instantaneous probability of the outbreak of war is constant over time.

L. F. Richardson (1956) notes that:

> This explanation of the occurrence of war is certainly far removed from such explanations as ordinarily appear in newspapers, including the protracted and critical negotiations, the inordinate ambition and the hideous perfidy of the opposing statesmen, and the suspect movements of their armed personnel. The two types of explanation are, however, not necessarily contradictory; they can be reconciled by saying that each can separately be true as far as it goes, but cannot be the whole truth. (p. 1258)

Does showing that the frequency of war outbreaks can be fitted with a Poisson distribution constitute an explanation at all, and if so, what exactly does it explain?

The Poisson distribution describes a surprisingly large number of natural phenomena. Kac (1983) gives an interesting illustration of the descriptive range of this construct, involving deaths of Prussian soldiers by horse kicks and the emission of alpha particles by a radioactive substance:

> The proportion of consecutive time intervals of duration τ (e.g., a week) during which k soldiers are killed by horse kicks is approximately $\exp(-a\tau)(a\tau)^k / k!$. Similarly, the proportion of consecutive time intervals of duration t (e.g., one second) during which k particles are emitted is approximately $\exp(-\alpha t)(\alpha t)^k / k!$ ... By a proper adjustment of units (amounting to setting $a\tau = \alpha t$) the two sets of data (one on soldiers killed, the other on alpha particles) will be difficult to distinguish. (p. 406)

There are many examples that could be given of natural phenomena—behavioral, social, and physical—certain characteristics of which are well described by statistical constructs. We have already noted the numerous statistical regularities discovered within the avalance of numbers produced by 19th-century counters and classifiers. To what extent should the identification of a statistical regularity be considered an explanation of the statistically regular event? More generally, what are we to make of statistical explanations? Are they really explanations? If they are not explanations, do they shed any light at all on the phenomena they fail to explain? If they are explanations, how do they relate to other types of explanations of the same phenomena? Are statistical explanations and causal explanations complementary, or are they qualitatively different and mutually exclusive ways of viewing the world?

Regression to the Mean

A widely recognized statistical phenomenon that is commonly referred to as an explanation is the phenomenon of regression, or, as Galton who first described it called it, "reversion," to the mean. It seems appropriate to begin a discussion of statistical explanation with a consideration of it.

The children of exceptionally tall parents are likely to be shorter, on average, than their parents, and the children of exceptionally short parents are likely to be taller, on average, than their parents. Children of parents with exceptionally high or exceptionally low intelligence are likely to have IQs closer to the average. Such facts are often attributed to chance and considered examples of the regression phenomenon of which Galton spoke.

In what sense does invocation of the concept of regression to the mean constitute an explanation of these and similar phenomena? Despite the fact that statistical explanation is usually distinguished from causal or physical explanation, terms such as "chance," "random process," and "probabilistic event" are sometimes given causal connotations. When regression to the mean is invoked for explanatory purposes, there is often the hint of an assumption of a force that draws the values of variables toward the mean much like the force of gravity draws small masses toward large ones. What the term really means is that one is likely to get a result when one samples (relatively) randomly from a population that is different in a specified way from the result one gets when one selects items from the same population that are known to be extreme or atypical in some way.

Consider a normally distributed variable, say person height. If one were to select several people from the end of this distribution representing the small fraction of unusually tall individuals and compare their heights with those of a random sample from the distribution, one would expect the heights of the randomly sampled group to be shorter on average—more representative of the distribution as a whole—than those of the people who were selected precisely because of their exceptional height. The regression-on-the-mean explanation of why especially tall parents tend to have children shorter than themselves rests on the assumption that the children are a more nearly random sample from the population than are the parents—by virtue of the fact that the parents, and not their children, were selected on the basis of their height—and are therefore likely to be more representative of the population as a whole. (The explanation does not require the assumption that the children are a totally random group; on that assumption, we would expect their average height to be very close to the population mean, not just closer to it than the mean of their parents' height.)

The notion of regression to the mean can be used to account for many phenomena. If one selects from all the mutual funds in the country the 10 best performers over an arbitrary 5-year period and then looks at the performance of the same 10 funds during the immediately following 5-year period, one will invariably discover that the performance of these funds was much closer to average during the second 5-year period than during the first one. I do not mean to suggest by this observation that the behavior of mutual fund managers is totally random, but a case can be made for the assumption that the degree to which fund performance is due to chance is much greater than the industry is likely to be willing to admit (Malkiel, 1985). As P. L. Bernstein (1996) puts it

> Over the long run, active investment managers—investors who purport to be stock-pickers and whose portfolios differ in composition from the market as a whole—seem to lag behind market indexes like the S&P 500 or even broader indexes like the Wilshire 5000 or the Russell 3000. Over the past decade, for example 78% of all actively managed equity funds underperformed the Vanguard Index 500 mutual fund, which tracks the unmanaged S&P 500 Composite; the data for earlier periods are not as clean, but the S&P has been a consistent winner over long periods of time. (p. 297)

Conjecture: If one compares the performance of a large number of mutual funds with that of the general stock market, as reflected in major indexes, over a random short period of time, say 1 to 5 years, one will find that a modest percentage—say 10% to 20%—of the funds outperformed the indexes during that period of time. And if one compares the same funds with the same indexes during the immediately following period of the same duration, one will again find that a modest percentage of the funds outperform the indexes. In both cases, if one sorts the funds into categories depending on how much they gained (or lost) during the period and plots the results as frequency distributions, one will find the distributions to be approximately normal (gaussian) with means differing from the means for the indexes by roughly the amount of what it costs the investor for transaction and fund management fees. And if one checks to see how the funds that outperformed the indexes during the first time period fared during the second period, one will find that they are scattered more of less randomly over the entire range of the second distribution. If one did the analysis here suggested—it may well have been done many times, but I cannot point to references—and one got the results I have imagined, one need not conclude that fund management is a completely chance affair, but it would be hard to avoid the conclusion that chance is at work to a nontrivial degree.

If one takes the batting averages of the 10 leading hitters among major-league baseball players during the first month of the season and compares those averages with the averages of the same 10 players during the final month of the season, one is highly likely to find the second set of averages to be closer to the league mean than the first. Similarly, if one compares the averages of the 10 leading players during the last month with the averages of the same players during the first month, one again will find the latter set to be closer to the league mean than the former. Each of these phenomena, among many others, can be described as regression to the mean.

Suppose that one claims to have a method by which one can increase children's IQs. To demonstrate the effectiveness of this method, its developer proposes to give a large number of school children an IQ test and then to use this method with the lowest scoring 10% of the group—the "hardest cases." Upon retesting after the intervention, the IQs of the students in this group prove to be higher on the average than when initially measured, and the intervention is declared a success. In fact, on the assumption that IQ measurements are unreliable to some degree, which is to say that the score produced by a given individual is likely to differ somewhat from one testing occasion to another, we would expect the scores to be higher on the second testing for the same reason that we expect exceptionally short parents to have children that are, on average, somewhat taller than themselves. It is the regression-to-the-mean phenomenon in another guise.

In general, regression to the mean can be expected whenever a set of values selected from one or the other end of a more-or-less normally distributed variable is compared with a set that can be assumed to be more representative of the entire distribution. The situation can be represented in the abstract in the following way. Suppose we were to have 30 people each roll a die five times (so the maximum score that one could receive would be five sixes or 30) and we identify the three people with the largest scores, the top 10% of our group, as the high-rollers. Now we repeat the process, have all 30 people again each roll the die five times. We would be quite surprised if all the high-rollers from the first pass retained their status on the second one. We would expect, assuming the die is fair, that their scores would be closer to average on the second pass than they were on the first—and indeed as likely to be below the mean as above it.

This illustration is a little extreme, because the cumulative score on five rolls of a die is presumably a totally chance affair. We would like to believe that performance in the financial world, on the baseball field, or in an intelligence-testing situation bears some relationship to ability and is not completely analogous to the tossing of a die. We can make the illustration more realistic by assuming, let us say, three types of dice, A, B, and C. A has the numbers 1 to 6 on its faces, B the numbers 2 to 7, and C, 3 to 8. Thus the maximum five-toss totals for the three dice are 30, 35, and 40, respectively. We divide our 30 subjects into three groups and have each group use a different one of the three dice. Now we have people who really differ in what we might think of as "native ability." Those who are stuck with die A will surely do worse, on average, than those who are fortunate enough to be rolling die C. Nevertheless, if we repeat the experiment described in the last paragraph, we are very unlikely to find that the three people who get the highest total scores on the first five rolls all remain in the top 10% on the second five rolls.

A number of researchers have shown how ignorance of the regression phenomenon can lead to the development or strengthening of unwarranted conclusions about the effects of various types of decisions. Imagine, for example a group of people whose performance with respect to some measure of merit is random; in particular, suppose that when performance is assessed for a specified period of time, the measures have a normal (gaussian) distribution. Now suppose that you, as the manager of this group reward (with praise or a bonus) the people who score in the top 10% of the range and punish (with criticism or a cut in pay) those who score in the bottom 10%. Suppose further that what you do has *no* effect on subsequent performance. By the principle of regression, we would expect that during the next performance period, the people who were rewarded will perform more poorly and those who were punished will do better. If you are not aware of this regression-based expectation, you may well conclude that the changes in performance resulted from your reward and punishment policy and come to the conclusion that punishment is an effective motivator whereas reward has the opposite effect.

Or suppose, focusing only on the bottom end of the performance scale, you decide after the first assessment to replace (or give special training to) the workers who scored in the bottom 10%. You will find, again assuming your actions have no effect, that the new (or now specially trained) workers will get higher scores, in general, in a subsequent assessment, a fact that you might, if you are unaware of the regression phenomenon, take as evidence of the effectiveness of your management decision (M. D. Cohen & Marsh, 1974). In most real-world situations it probably would not be reasonable to assume that performance was due totally to chance; it is not unreasonable, however, to assume that chance plays some role and that regression to the mean therefore should not be completely ignored.

Statistical Versus Causal Explanations

> Inexperienced researchers and laypeople alike usually overestimate the influence of systematic factors relative to chance factors. (Abelson, 1995, p. 7)

Regression to the mean will be satisfactory as a complete explanation only when the distribution of the measures of interest around the mean can be assumed to be due completely to chance. This is an assumption that fund managers, at least, are not likely to be pleased to make with respect to mutual fund performance. Baseball players, at least those with high batting averages, probably are not more willing to make it with respect to batting performance. One can, of course, believe that the value of a variable is due in part to chance and in part to other factors. So it is not necessarily unreasonable to look for deterministic factors that may be operative even when it may be assumed that there is reason to believe that chance is having some effect. And, conversely, chance can be an influential factor even when there is reason to doubt that it is the only factor at work.

Major-league baseball players who win rookie of the year almost invariably perform less well the year following that of their award (Nisbett, Krantz, Jepson, & Kunda, 1983). On the assumption that chance plays some role in determining the kind of performance that leads to selection of rookie of the year, this may be viewed as an instance of regression to the mean. On the other hand, it seems ungenerous, if not unreasonable, to assume that that performance is totally due to chance. That being so, it is not unreasonable to attempt to understand, in individual cases, precisely how performance changed and to identify, when possible, nonchance factors that may have contributed to the change for the worse.

Statistical and causal explanations are not necessarily mutually exclusive. It is not impossible that one might be able to account for the same set of phenomena in both ways and that the accounts would not be contradictory. When this is possible the causal explanation is likely to be a deeper explanation than the statistical one. For example, one may attribute to chance the fact that roughly 50% of a large number of tosses of a fair coin come up heads, but if one knew enough about the physics of the individual tosses, one might be able to account for them on a toss-by-toss basis, without resorting to the concept of chance. (Whether accounting for each individual outcome would constitute also an explanation of why the proportion of heads in a large number of tosses is close to .5 is another question.) Often, though not always, the need to resort to probabilistic explanations reflects a lack of information that is, in principle, obtainable and that would provide the basis for a causal explanation. Statistical explanations should be resorted to primarily when deterministic causal explanations are either impractical or, as in the domain of particle physics, perhaps impossible.

This is not to deny that people sometimes look for causal explanations when it may be inappropriate to do so. One of the conclusions that Tversky and Kahneman (1971) have drawn from studies of experimental psychologists' intuitions about appropriate sample sizes in psychological experiments is that experimenters rarely attribute a deviation from expectation in their results to sampling variability; instead, they find causal explanations for the discrepancies they observe. And because the experimenters tend to be satisfied with the explanations they generate, they preclude themselves from having the opportunity to recognize sampling variation in action and consequently inappropriate belief in "the law of small numbers" is not disconfirmed.

At the beginning of the discussion of the concept of regression to the mean, I suggested that terms such as "chance," "random process," and "probabilistic event" are sometimes given causal connotations. And in the ensuing comments I more than once alluded to chance as a causative agent. Exactly what might it mean for something to be *determined* by chance? It sounds as though one is imputing to "chance," whatever that is, the power to determine the outcomes of events—not the outcomes of individual events, to be sure, but the aggregate outcome of multiple events. Chance is being identified as the *cause* of the aggregate outcome. This makes chance sound very deterministic, which appears to be a contradiction in terms.

One might say that this is not correct, that all one is doing is *describing* the outcome, not accounting for it, when one attributes it to chance. But Hacking (1987b) notes that at one point in the evolution of statistical thinking, "it was an essential part of the doctrine of chances that there was always an underlying causal structure. It was the task of the analyst to find that structure," but that at a later point, "people became indifferent to that" (p. 53). According to the later view, statistical law became autonomous, which is to say usable "to explain something else, without itself having to be reduced" (p. 53). This appears to make the question of what accounts for the predictability of chance events—the lawfulness of chance—moot. I do not find this completely satisfying. Chance, in my view, is a great mystery; I doubt if anyone understands it very well.

Krüger (1987b) also points out that acceptance of probability as an explanatory construct was slow:

> Although in the course of the nineteenth century, statistical practice and a corresponding amount of probability theory spread rapidly through various disciplines, most scientists and philosophers remained opposed to taking probability as fundamental or irreducible, or according it an explanatory function. Indeed, that function appeared to imply the reality of possibilities or of indeterminateness; hence to recognize it seemed to involve too high an ontological price. At any rate, there is a time lag between the widespread use of statistics and probability on the one hand and the adoption of a probabilistic view of, or attitude toward, reality (or parts of reality) on the other. (p. 60)

No less a figure than Kant insisted that nothing happens through blind chance, which implies that all phenomena have deterministic explanations even if one cannot discover what they are. According to this view, which dates back at least to the classical Greek atomists (e.g., Leucippus and Democritus), "chance is excluded not only from the domain of scientific concepts but also from the world of events; laws of chance or probability theory cannot possibly refer to reality" (p. 62). Krüger points out that many 19th-century scientists and philosophers considered it obvious that statistical regularities must have deterministic explanations. He credits Maxwell with transforming what had been descriptive in social science and statistics into explanatory accounts of observable phenomena like heat, flow, and diffusion.

That knowledgeable people can interpret its significance in quite different ways is illustrated by the contrasting views of Monod (1972) and Polkinghorne (1986). Monad takes the operation of chance in the processes of the world as evidence of their meaninglessness. Polkinghorne views it as providing an insight into the design of the world:

> When I read Monad's book I was greatly excited by the scientific picture it presented. Instead of seeing chance as an indication of the purposelessness and futility of the world, I was deeply moved by the thought of the astonishing fruitfulness it revealed inherent in the laws of atomic physics ... the fact that they have such remarkable consequences as you and me speaks of the amazing potentiality contained in their structure. From this point of view the action of chance is to explore and realize that inherent fruitfulness. (p. 54)

The "laws of chance" is the name we give to the fact that certain events that are irregular and unpredictable individually are regular and predictable in the aggregate. As an explanatory construct, chance, or the idea of the laws of chance, raises the question of what an explanation is. What do we mean when we say that the outcome of the toss of a coin is determined by chance? Or that the laws of chance dictate that if a fair coin is tossed a large number of times, it will come up heads on about half of the tosses? In what sense have we explained anything by such claims?

Perhaps we need to recognize levels of explanation. A concept that is used as an explanatory construct at one level must be explained, if at all, at a deeper level. I say "if at all" because instead of attempting to explain a concept, one might opt to take it as a given—as a primitive of one's explanatory system. One might elect, for example, to treat chance as such a construct—take it as a given and decline to attempt to explain it in terms of more basic concepts. This does not preclude one from studying *how* it works; it simply admits an inability, or at least a disinclination to try, to determine *why* it works.

So if we accept the laws of chance, much as we accept the law of gravity, as descriptive of a characteristic of the way things are, and let it go at that, we may invoke the concept to account for events that we cannot account for deterministically. We can, in other words, attribute to chance events or outcomes that we cannot account for in a more traditional cause–effect way. But we should recognize, it seems to me, that this is explanation at a less than fundamental level. By defining random sampling as sampling in such a way that every member of the population has an equal likelihood of being in the sample, we can invoke the concept of regression to the mean, for example, to account for a variety of phenomena. Such an explanation accounts very well for the phenomena, given the definition of random sampling, but it does not explain why random sampling works.

Identification of Underlying Stochastic Processes

When a variable is distributed in some regular fashion—say in a way describable by a simple mathematical function—one is prompted to look for some process that would produce a variable with values matching that distribution. And when one discovers two or more variables with the same or similar distributions, one naturally wonders whether the values of those variables are determined by the same or similar underlying processes. The search for stochastic processes that will yield values of variables that match specific distributions has proved to be a fruitful line of investigation in probability theory and statistics. The idea is illustrated by reference to a few simple processes that will produce specific distributions.

The variable of interest in the first example is the number of heads in 20 tosses of a fair coin. This variable can take on any value from 0 to 20, inclusive. Suppose we define an event as 20 tosses of the coin and an outcome as the number of heads obtained in that event. Imagine performing a rather tedious experiment in which we record the outcomes of a large number—say 10,000—of such events. If we performed this experiment and graphed the results, we would have a distribution of event outcomes very close to that shown in Fig. 7.1, which can be described mathematically as a binomial distribution.

The second example involves rolling a fair die, which, by definition is equally likely to turn up any of its faces. If we were to roll such a die a very large number of times, we would find each of the values 1 through 6 coming up roughly equally often. That means, if we were to plot the distribution of values it would be approximately rectangular. Now suppose we rolled a *pair* of fair dice 10,000 times. What would we get as a distribution of *sums*? This distribution, which would be a distribution of the values 2 through 12, would look like that shown in Fig. 7.2.

There would be about twice as many 3's as 2's, because there is only one way to roll a 2 (1 on each die) whereas there are two ways to roll a 3 (a 1 on the first die and a 2 on the second or a 2 on the first die and a 1 on the second). There would be about three times as many 4's as 2's because there are three ways to roll a 4: 1-3, 2-2, 3-1. And so forth. We can extend this thought experiment indefinitely by imagining rolling three, four, or any number of fair dice at a time. What we would produce as a distribution of sums in any instance is easily determined by figuring the number of combinations that will produce each possible sum and the number of ways (permutations) in which one could get each such combination.

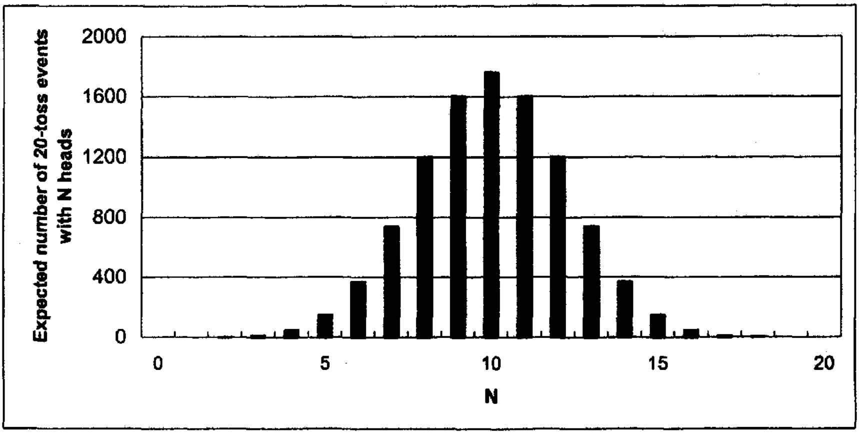


FIG. 7.1. Expected number of events that would have *N* heads in 10,000 events in each of which a coin is tossed 20 times.

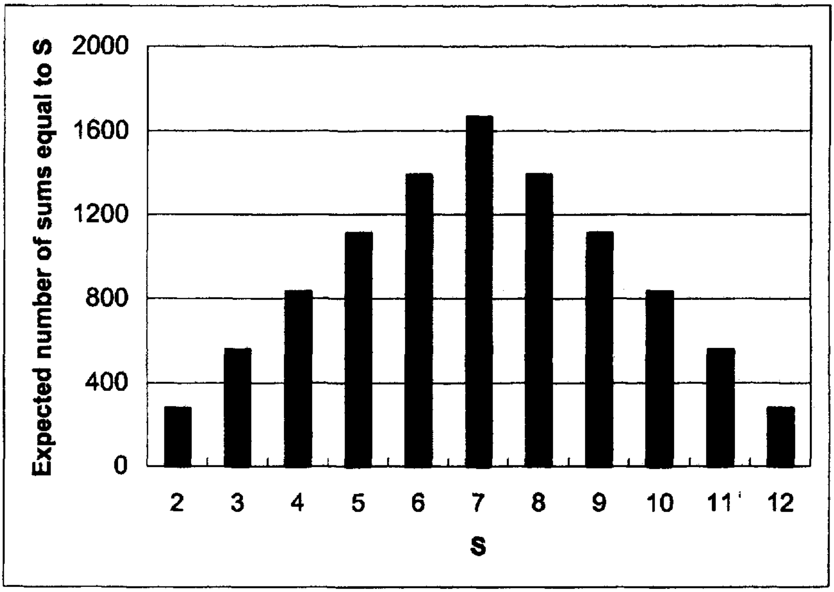


FIG. 7.2. Expected number of sums equal to *S* in 10,000 tosses of a pair of dice.

The final example involves another coin-tossing exercise. In this case suppose we defined an event as the number of successive heads tossed before the occurrence of a tail. That is, on each trial we would toss the coin until it came up tail, and the value of the variable for that trial would be the number of heads tossed; the occurrence of a tail would terminate the trial. Obviously, this variable can take on any integer value equal to or greater than 0. When we plotted the distribution of values obtained in 10,000 trials we would expect to get a distribution approximately lie that shown in Fig. 7.3 which can be described by the exponential function

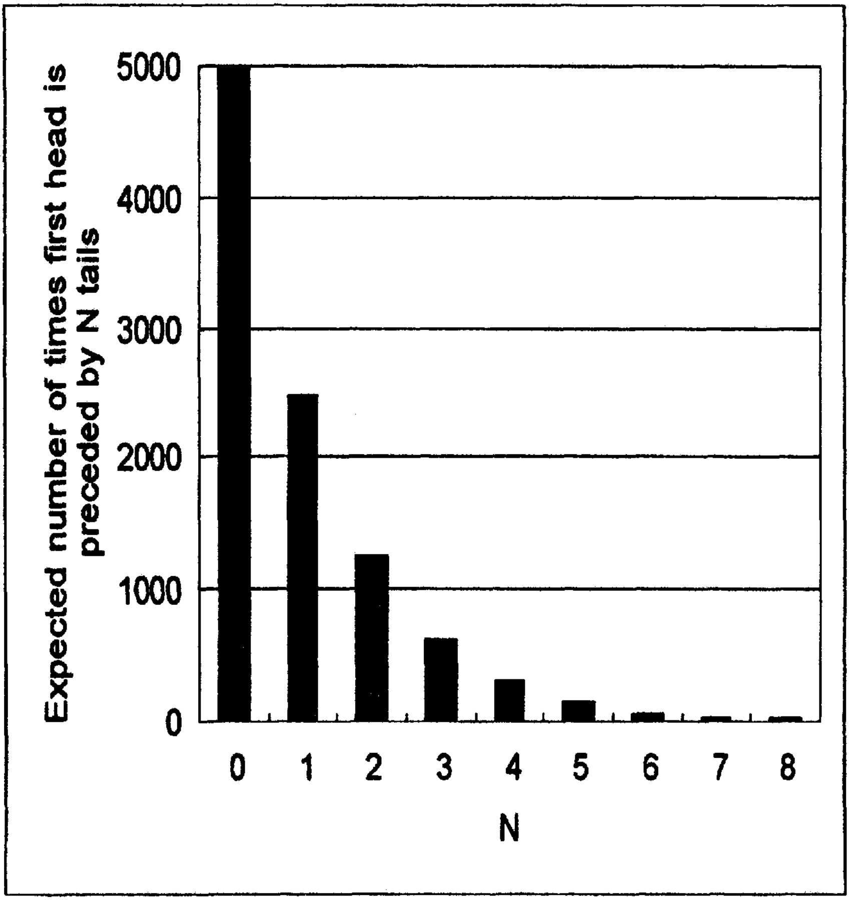


FIG. 7.3. Expected number of times the first head will be preceded by *N* tails in 10,000 trials.

$$f(x) = \frac{n}{2^{x+1}},$$

where n, in this case, equals 10,000.

Given an understanding of how the aforementioned distributions are generated, their shapes make intuitive sense. We would expect, before doing the first coin-tossing experiment, for example, to end up with relatively many outcomes with approximately equal numbers of heads and tails, and relatively few with nearly all heads or nearly all tails. Similarly with the other experiments, understanding the process by which the distributions are generated leads us to expect the outcomes we would get. Why we should have these expectations, and why they should be borne out, are interesting questions to which I do not think we have the answers.

Many variables in nature are distributed in regular ways and sometimes it is possible to describe what appear to be simple processes that will produce the same distributions. Does the identification of such a process constitute an explanation of why a natural distribution is what it is? Showing that a particular distribution could have been produced by a specified process is not equivalent to showing that it was produced by that process. Moreover, when a distribution can be assumed to have been produced by a specified process, the question remains as to why the process has the properties it does.

Statistics in the Service of Argument

> The purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric. (Abelson, 1995, p. xiii)
>
> Good statistics involves principled argument that conveys an interesting and credible point. (Abelson, 1995, p. 1)

The idea that the results of statistical tests should be used primarily as aids to human judgment, and seldom if ever as definitive justification for accepting or rejecting a hypothesis has been defended forcefully by Abelson (1995) in his presentation of "statistics as principled argument." Abelson identifies five factors that, in his view, determine the persuasive force of a statistically supported argument, and organizes them around the acronym MAGIC:

- Magnitude: "The strength of a statistical argument is enhanced in accord with the quantitative magnitude for its qualitative claim" (p. 12). Larger effects are generally more persuasive than smaller ones, which is not to say that small effects are never interesting or important. Abelson introduces the idea of "causal efficacy," which he defines as effect size divided by "cause size," and according to which, "A large effect from a small variation in the cause is the most impressive, whereas a small effect arising from an apparently large causal manipulation is the most anticlimactic and disappointing" (p. 48).

- Articulation: "the degree of comprehensible detail in which the conclusions are phrased" (p. 12). A more detailed conclusion (A > B > C) is more persuasive than a less detailed one (A, B, and C differ).
- Generality: "the breadth of applicability of the conclusions" (p. 12). Broad conclusions, which are likely to require for support data from several related studies, are more persuasive than narrow conclusions, like those to which single studies are usually limited.
- Interestingness: "for a statistical story to be *theoretically* interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue" (p. 13).
- Credibility: "the believability of a research claim" (p. 13). Credibility, Abelson suggests, depends on both the soundness of methodology and theoretical coherence of the claim.

Abelson's (1995) book is a compelling elaboration of how statistics can be used effectively to support judgment and reasoned argument, with emphasis throughout on the principles noted previously. It documents numerous common misconceptions about statistical testing and many ways in which statistics can be, and have been, used inappropriately and to the detriment of psychological research. It is crammed with clear, practical, and often engagingly witty advice to statistics users, novices and experts alike: "*Never flout a convention just once.* In other words, either stick consistently to conventional procedures, or better, violate convention in a coherent way if informed consideration provides good reason for so doing" (p. 70). "Don't be overjoyed when your test statistics come out whopping. Be suspicious" (p. 90). "Omnibus testing is like playing the guitar with mittens on" (p. 105). "A wise general practice in the statistical treatment of complex data arrays is first to display them graphically , and do rough, simple quantitative analyses. These will give a feel for the potential meaning of the results; only then should you resort to complex refinements" (p. 128). "It is a good idea for researchers themselves to conduct one or two replications before getting too carried away by the force of their initial claims" (p. 133). "One might say that isolated claims are not *robust.* Investigators who feel that their results march with full generality into the annals of science are kidding themselves" (p. 149). "To be interesting, a result has to make you think about the topic—or at least make you want to think" (p. 160). "Research claims are regarded as guilty of obvious artifactual possibilities unless these are explicitly and adequately dealt with" (p. 180).

THEORIES OF STATISTICAL INFERENCE

Although some seminal suggestions regarding the use of statistics for inferential purposes had been put forth as early as the middle of the 19th century—for example, the work of Gustav Radicke (Coleman, 1987)—what might be referred to appropriately as theories of statistical inference were not developed until the 20th century. Gigerenzer and Murray (1987) point out that something of an "inference revolution" occurred in psychology between 1940 and 1955, as a consequence of which "inferential statistics" became indispensable for psychologists. Unfortunately, as the same authors also note, many textbooks that are used to teach students techniques of statistical analysis and experimental design present the material as though the theoretical underpinnings of statistical inference were noncontroversial. Often theory, let alone theoretical controversy, is not even mentioned. Algorithmic procedures are given for calculating various statistical quantities and for making tests of statistical significance; follow the computational recipe, consult the appropriate tables, report the results, and rest easy that science has been served. In fact, the development of the theoretical underpinnings of statistical inference was full of controversy and differences of opinion persist regarding some of the foundational issues.

Many people contributed to the development of mathematical statistics during the early part of the 20th century. One who deserves special mention is Karl Pearson. He is remembered primarily for the correlation coefficient that bears his name, but his influence went way beyond this contribution. He advanced the concept of a skew distribution, identified certain measures (mean, standard deviation, symmetry, and kurtosis) as the important parameters of a distribution, created the first goodness-of-fit test using the chi-square statistic, and established—along with Francis Galton and Raphael Weldon—*Biometrika,* which published primarily distribution measurements on biological variables motivated by an interest in Darwin's then new theory of evolution. Pearson is also remembered for the long-lasting feud between him and another major figure in the statistical world of the early 20th century, Ronald A. Fisher; the jury is still out on the question of who was the more contentious of the two.

Three different schools of thought have defined much of the controversy around statistics during the 20th century; one derives from the work of R. A. Fisher, another from that of Jerzy Neyman and Egon Pearson (Karl Pearson's son), and a third from the approach originally put forth by Thomas Bayes. Gigerenzer et al. (1989; Gigerenzer & Murray, 1987) see each of these views as "a considerable advance" over earlier ones, but they note that none of the three has emerged as the clear winner, and that the issues on which they are divided are deep and fundamental: "The different schools often disagree fiercely about basic issues, and value-laden words from ordinary speech such as 'efficient,' 'unbiased,' and 'coherent,' have been enlisted as names of central concepts in the various theories. By implication, rival approaches are charged with inefficiency, bias, and incoherence" (p. 90). The rivalry, sometimes acrimonious, among the pioneers, especially between Fisher and Karl Pearson, has been described in some detail by Salsburg (2001). Salsburg's book recounts also the contributions of other notables to the development of statistics as a mathematical discipline.

The approaches of each of these schools are widely applied, those of Fisher and Neyman–Pearson primarily in the experimental sciences and those of Bayes more in economics and related disciplines. Fisherian and Neyman–Pearson ideas about statistical inference have been extremely influential in the development of methodology and especially conventions of experimental design in psychology. The Bayesian approach has been studied widely by psychologists as a normative model of everyday reasoning and decision making and has motivated a great deal of experimentation designed to determine the extent to which people naturally behave as Bayesians. The Fisherian and Neyman–Pearson views are considered briefly in the next two subsections; the Bayesian approach has already been discussed in chapter 4. Curiously, Bayesian analysis has not been applied much to the interpretation of experimental data, even by those who consider it a normative model of reasoning and decision making.

Fisherian Inference

The Fisherian approach, which was set forth in *Statistical Methods for Research Workers,* first published in 1925, centers on designing experiments to test the *null hypothesis,* the hypothesis that two samples were randomly drawn from the same population. The statistical question that Fisherian significance tests, such as the analysis of variance, are intended to answer is the question of the probability of obtaining a difference between observed and hypothesized data of a specified magnitude by chance. By convention, when the probability of obtaining an observed difference by chance is determined to be less than a specified criterion, say .05, the null hypothesis—the hypothesis that there is no nonchance difference—is said to be rejected at a level of confidence defined by that criterion. As already noted, the logic of the significance test is not universally accepted.

Gigerenzer et al. (1989) distinguish between a *substantive* null hypothesis and a *statistical* null hypothesis, and argue that it is only the latter that can be rejected by a Fisherian significance test. The substantive null hypothesis is the hypothesis that an experimental treatment has had no effect; the statistical null hypothesis is the hypothesis that two samples have been drawn from the same population. Systematic errors in the design of an experiment could lead either to the rejection of the statistical null hypothesis when the substantive null hypothesis is true or to failure to reject the statistical null hypothesis when the substantive one is false. For this reason, proceeding inferentially from the rejection of a statistical null hypothesis to the acceptance of the hypothesis of a causal effect, a common practice in the interpretation of the results of psychological experiments, is not as straightforward as typically assumed (Gigerenzer & Murray, 1987).

It is conventional in the Fisherian approach to set a strict criterion of rejecting the null hypothesis; .05 is commonly used, as is .01. This means that, given the assumptions on which the test rests, the probability of rejecting the null hypothesis if it is really true is small (less than .05 or less than .01 for these two criteria). It means also, however, that the probability of failing to reject the null hypothesis if it is really false is likely to be quite large. (Fisherians never speak of accepting the null hypothesis; the only two options one has are to reject it or to fail to reject it.) The rationale for putting this kind of bias in the system is the assumption that, in the context of scientific research of the kind for which statistical hypothesis testing is usually done, failing to reject the null hypothesis when it is really false is a less objectionable type of error than rejecting the null hypothesis when it is really true. This bias is an analogue of the preference in law for failing to convict a guilty party over convicting an innocent one. Fisher did not put as much stock in a single experiment that yielded a $p < .05$ than have many of the subsequent users of his ideas; he stressed the importance of replication arguing that a phenomenon should be considered experimentally demonstrable when one knows how to design an experiment so that it will almost always yield a statistically significant result.

In a defense of the Fisherian approach to statistical inference, Macdonald (1997b) notes that:

> A low significance level does not require one to accept the presence of an effect. One's acceptance can be influenced by such factors as the effect's importance, its consistency with previous findings, its compatibility with one's existing beliefs and one's confidence in the researchers and in the study's methodology. Indeed this acceptance can depend on aspects of the data of which the test takes no account (e.g., inconsistencies within the data, overly good fits, unexplained peculiarities, errors in other parts of the analyses, etc.). (p. 339)

Macdonald argues too, that, although Fisher rarely spoke of power—the probability of rejecting the null hypothesis conditional on its being false—in his writings, it does not follow that the notion is incompatible with the approach.

Fisher himself appears to have been a colorful, strong-willed, brilliant individual and his work probably had at least as much influence on the development of inferential statistics as that of any other individual. His interests were not confined to statistics and experimental design; he wrote extensively and influentially on genetics as well (Salsburg, 2001; J. R. Turner, 1987), sandwiching his major work on this topic, *The Genetical Theory of Natural Selection,* published in 1930, between *Statistical Methods for Research Workers,* 1925, and *The Design of Experiments,* 1935. His long-running feuds first with Karl Pearson and later with Egon Pearson and Jerzy Newman are well documented (Box, 1978; Salsburg, 2001), although seldom discussed in statistics textbooks.

Neyman–Pearson Inference

Jerzy Neyman was 37 years younger than Karl Pearson and a contemporary of Karl's son, Egon. The Pearson of the Neyman–Pearson collaboration was Egon. Neyman and E. Pearson's (1928a, 1928b, 1933) approach to statistical inference can be seen as, in part, a reaction against what its proponents perceived to be the overly one-sidedness of the Fisherian method of hypothesis testing. Instead of posing questions that could be answered only in a yes–no fashion—yes the null hypothesis is rejected, no it is not—the Neyman–Pearson approach provides for the consideration of how the weight of evidence contributes to the relative statistical plausibility of each of two competing hypotheses, one of which can be the null hypothesis.

Use of this approach involves not only articulating the hypotheses that are to be considered but also specifying the criterion that will be used to make a selection between them. One selects a decision criterion on the basis of the relative benefits and costs of the various possible ways to be right or wrong—deciding in favor of Hypothesis 1 when Hypothesis 2 is true, and so on. Newyman and Pearson saw their approach to hypothesis testing as one among other alternatives and stressed the need for users of statistical inference techniques to exercise judgment regarding the appropriateness of specific techniques to their own situations.

In somewhat oversimplified terms, Neyman–Pearson hypothesis testing can be equated with either–or decision making, whereas the Fisherian approach is better described as yes–no. Because of the either–or nature of the former, when one of the hypotheses is the null it is possible to compute the probability of both types of error and thus also power. So, unlike with Fisherian testing, acceptance of the null hypothesis is a legitimate outcome in this case.

Macdonald (1997b) describes both the Fisherian and Neyman–Pearson approaches to statistical inference as being "concerned with establishing that an observed effect could not plausibly be accounted for by sampling error" (p. 334). When an achieved significance level—the probability of obtaining an effect as extreme as the one observed—is sufficiently low, the hypothesis that the obtained effect could plausibly be due to sampling error is rejected, and rejection of this hypothesis entails also rejection of the hypothesis of a true effect opposite in direction to that observed.

Bayesian Inference

The Bayesian approach to hypothesis evaluation was discussed at length in chapter 4. Suffice it to say here that this approach has not been applied widely by psychologists to the analysis of the results of their experimental studies, despite arguments advanced by many writers regarding the advantages of such analyses (Edwards et al., 1963; Gelman et al., 1995; Good, 1983a; Greenwald, 1975; Lindley, 1984; Rindskopf, 1997; Rouanet, 1996; Rozeboom, 1960; Rubin, 1978).

Hybrid Inferential Statistics

Gigerenzer et al. (1989) review the debate between the followers of Fisher and those of Neyman–Pearson (as well as those of Bayes) in some detail and note that it has not ended. They note too the remarkable fact that little hint of the historical and ongoing controversy is to be found in textbooks that are used to teach statistical significance testing to its potential users. The amalgamation of ideas from different schools of thought that is commonly taught fails to do justice to the richness and complexity of their philosophical underpinnings and promotes an uncritical and doctrinaire application of certain types of statistical analyses as the *sine qua non* of good psychological science:

> The need for personal judgment—for Fisher in the choice of model and test statistic; for Neyman and Pearson in the choice of a class of hypotheses and a rejection region; for the Bayesians in the choice of a prior probability—as well as the existence of alternative statistical conceptions, were ignored by most textbooks. As a consequence, scientific researchers in many fields learned to apply statistical tests in a quasi-mechanical way, without giving adequate attention to what questions these numerical procedures really answer. (p. 106)

One must wonder why this state of affairs exists. Who has decided that students, especially students who are preparing to do scientific research, have no need to know about the philosophical disputes that have raged regarding the justifiability of some of the fundamental tools that they are learning how to use? Are students being "protected" from these controversies? Are instructors concerned that exposure to them would undermine students' confidence in the efficacy of statistical inference? Are instructors reluctant to try to make sense of them? Are they themselves unaware of them? Are Gigerenzer and Murray (1987) correct in attributing the widespread belief among psychologists in the "illusion" of a single statistical theory for valid inference from data to hypotheses in part to self deception rooted in the strong desire to have such a theory? "Psychologists seem not to wish to be cured of these illusions, for the consequence would be the abandonment of the indispensable instrument and hence the abandonment of the unification of psychological methodology institutionalized by the inference revolution" (p. 25).

Gigerenzer et al. (1989) also note that significance testing, the rules of which are based on an amalgam of the ideas of Fisher and Neyman–Pearson that neither Fisher nor Neyman and Pearson would be likely to endorse, has become almost the only statistical tool that is used in sociology and psychology, and that in these fields other tools, such as confidence intervals, the likelihood function, and Bayesian inference, have been given little attention by comparison. They note too the strange incongruity in the fact that many researchers who have done experimental studies of human reasoning under uncertainty have concluded that their subjects are irrational because they do not reason in accordance with Bayes' theorem, whereas they themselves use Fisherian or Neyman–Pearsonian statistics or a mixture of them to test their own hypotheses.

As institutionalized in university curricula and journal editorial policies, the hybrid theory, as Gigerenzer et al. (1989) call the amalgam, with its sharp focus on statistical significance testing has been a major determinant of how sociological and psychological research is done: "In some fields, a strikingly narrow understanding of statistical significance made a significant result seem to be the ultimate purpose of research, and non-significance the sign of a badly conducted experiment—hence with almost no chance of publication" (p. 108). Support for this view comes from a study by Sterling (1959) that showed that as of 1955 more than 80% of the articles in several leading psychology journals used significance tests to justify conclusions from the data. Gigerenzer et al. claim that as of the late 1980s the figure was somewhere between 90% and 100%.

The point of this discussion is not to deny the usefulness of statistical significance testing as a research tool. Used thoughtfully with an appreciation of its limitations, it can be a powerful aid to extracting information from noisy data. Used unthinkingly without an understanding of the assumptions on which various tests rest and of the controversies that have surrounded some of those assumptions, it can become a fetish and an impediment to scientific advance. Statistics texts typically do make a distinction between statistical and theoretical or practical significance, but the dominance of statistical significance testing as an experimental paradigm has obscured this distinction in practice. The purpose of research is to further our understanding of how the world works, not just to discover statistically significant differences, and only to the extent that the discovery of such differences really serves the larger purpose does it contribute anything of worth to the enterprise.

Noting that early psychologists took physics as their model for an experimental science, Gigerenzer (1987) raises the interesting question of why they did not follow the lead of quantum theory, and proposes the answer that "quantum theory seemed to violate two ideals connected with the struggle for *certain* knowledge: *determinism* and *objectivity*" (p. 12). Probability, he argues, was pressed into the service of these ideals through the way in which statistics was applied to the drawing of inferences from experimental data: "Probabilistic thinking was used as a means toward objectivity in the classical sense of separating the experimenter from his knowledge. Such was the role of inferential statistics as a mechanization of the experimenter's inference from data to hypothesis" (p. 12).

The amalgamation of the ideas of Fisher and Neyman–Pearson into the hybrid form of statistics that is generally taught in psychological experimental design texts is seen as integral to maintaining the objectivity of psychology as an experimental science:

> The connection I shall draw between the kind of inferential statistics established in psychology and objectivity is based on the following observations: (1) There was *a single dominant theory* of inferential statistics that (2) was taught *anonymously* (i.e., without indication of its multiple and sometimes contradictory sources) as "truth" per se. (3) Problems stemming in part from the fact that the theory was spliced together from theses different sources went *unacknowledged,* (4) alternative theories were *neglected,* and (5) this dominant hybrid theory was *institutionalized* by editors of journals and internalized by authors as the one "true" path to experimental knowledge (and therefore toward publication). (Gigerenzer, 1987, p. 12)

Determinism could coexist with probabilistic thinking during the first half of the 20th century, Gigerenzer argues, because "probabilistic thinking was enlisted in the service of determinism" (p. 15). The appearance of objectivity was promoted by eliminating the need for the experimenter's judgment through the application of mechanical statistical procedures to the interpretation of data; statistics became "a means for the mechanization of inductive inference" (p. 25). Neglecting to acknowledge the controversial history of statistics and the existence of approaches alternative to the one that was almost universally taught were essential to the maintenance of this appearance, in Gigerenzer's view.

The amalgam, or hybrid theory, could of course be exactly right. The fact that neither Fischer nor Neyman and Pearson would own it as his, or their, own does not necessarily make it wrong. Indeed, given the sharp differences of opinion between Fisher on the one hand and Neyman and Pearson on the other regarding how statistical inferences should be made, it would not be surprising if a theory that stood the test of time turned out to show the influence of both views but to not be totally consistent with either. A theory of statistical inference should be judged on its merits and not on its historical pedigree. Critics of the hybrid view generally would argue, I believe, that not only is this view not completely consistent with Fisher's theory or that of Neyman and Pearson, but that it also does not have a self-consistent theoretical foundation of its own.

Signal Detection Theory

A strong argument could be made for the idea that the single most important insight that has ever been expressed about decision making is the recognition that in most choice situations there are more ways than one to be wrong not all of which are equally palatable to the decision maker. Perhaps the most well-known illustration of this principle was Pascal's famous wager, which goes roughly as follows: God exists or He does not. Reason cannot decide the issue. One must wager. One can bet that God exists or that He does not, and in both cases one can be right or wrong. Suppose that one bets that God exists. If one is right, one gains infinitely much; if one is wrong one loses little. Suppose that one bets that God does not exist. If one is right, one gains little; if wrong, one loses infinitely much. Clearly, Pascal argued, one should bet that God exists.

It is doubtful if any other wager, actual or proposed, has generated as much commentary and debate as this one. Many variants of Pascal's wager, in which the possible outcomes are not so disparate as those expressed by him but that still represent greatly differing levels of desirability, have been constructed since Pascal's time. The point of all of them is to demonstrate that the way one wishes to "bet" on an uncertain outcome is likely to be determined not only by how likely one believes the various thinkable outcomes to be but also on the anticipated satisfaction or regret associated with betting correctly or incorrectly (selecting an outcome that does not occur) in the various possible ways. The wager illustrates in an intuitively compelling way that there can be more to rationality in uncertain choice situations than simply making selections so as to maximize the probability of being correct.

This insight has had considerable influence in experimental psychology through the application of the statistical theory of signal detection (D. M. Green & Swets, 1966; Swets, 1964; Tanner & Swets, 1954). In its basic form the theory deals with the problem of detecting signals in noise; the observer's (detector's) task is to say whether an observation interval contains only noise or a signal as well. Fundamental to the theory is a sharp distinction between the sensistivity of the observer and the criterion that the observer uses to determine whether to report that a signal is present.

The situations to which signal detection theory was originally applied were designed so that an observer could not increase the probability of reporting the presence of a signal when one had actually occurred without also increasing (by a different amount) the probability of reporting the presence of a signal when one had not occurred. In order to attain a desired probability of reporting a signal when it was really present (the probability of a "hit"), one had to settle for accepting also some nonzero probability of reporting a signal when it was not present (the probability of a "false alarm").

Such situations are very common in life; in order to capture more of the real signals, one has to be willing to pick up also more false alarms. Or to look at it from another perspective, the decision problem becomes that of deciding what kind of a compromise one is willing to strike between failing to capture real signals on the one hand and incorrectly counting nonsignals as signals on the other. A "signal," as the word is being used here, could be a literal signal, as in the context of a sonar operator trying to decide whether a signal emitted by a ship has occurred in a background of acoustic noise, or it could be a figurative one, as when we treat the outcome of a diagnostic test as a possible signal for the presence of a specific disease.

The problem of the covariation of hits and false alarms is illustrated clearly in the case of medical diagnosis. Many disease symptoms are continuous in nature—a high temperature, for example, is often symptomatic of illness, but temperature can vary continuously over a considerable range and it is impossible to select a point on this continuum such that all higher temperatures occur only in the presence of disease and no lower ones ever do. Thus if one uses temperature as an indication of disease, wherever one places the divide between "normal" and "abnormally high" temperatures, there are likely to be some healthy people with a temperature above this point and some sick ones with a temperature below it. (For purposes of this illustration, I am ignoring that an abnormally low body temperature also can indicate a medical problem.) In order to decrease the percentage of healthy people whose temperature is above the "abnormally high" criterion, we can raise that criterion, but in so doing we have to accept the fact that we will also increase the percentage of sick people whose temperature reading will fall below it; conversely to decrease the percentage of sick people whose temperature will fall below the criterion, we can lower it, but in doing this we also increase the percentage of healthy people whose temperatures will be classified as abnormally high.

The problem is very common and because signal detection theory provides a quantitative means of dealing with it, the theory has found application in a great many contexts (Swets, 1986; Swets, Dawes, & Monahan, 2000a, 2000b). By quantifying what can be expected given what is known or assumed about the nature of the "signal" that one is trying to detect and the sensitivity of the detection process, it can help one decide on a decision criterion that takes into account both of the types of mistakes that can be made. In the case of the medical illustration, if temperature were the only symptom one had on which to base a diagnosis of a particular disease (which of course it is not), where one would set the criterion for "abnormally high" would depend in part on the relative seriousness of the two possible kinds of mistake—judging a healthy person to be sick and judging a sick one to be healthy.

AMBIGUITY IN STATISTICAL ASSERTIONS

The word *ambiguity* sometimes has a specific technical meaning in the context of decision theory, which is illustrated by a comparison of the two following situations. Situation 1: You are to draw a marble from a bag that you know contains an equal number of red and blue marbles. Situation 2: You are to draw a marble from a bag that you know contains either all red marbles or all blue marbles, but you have no reason to suspect that either possibility is more likely than the other. In either case if asked the probability of drawing a red marble, you would undoubtedly say .5. And if faced with the necessity of making some important decision on the basis of the color of the marble you got on a single draw, it would seem that you should be equally willing to draw from either bag, inasmuch as the degree of uncertainty regarding the outcome of a draw is the same in both cases.

Nevertheless, the situations are not identical. In the first case, you know the proportion of marbles of each color in the bag; in the second case, you do not. As it turns out, people are not always indifferent to these two situations. In the jargon of decision theory, the second situation is said to be not only uncertain but ambiguous, the term ambiguity having come to represent, in this context, a specific type of uncertainty (Einhorn & Hogarth, 1985; Ellsberg, 1961; Frisch & Baron, 1988; Gardenfors & Sahlin, 1983). People often show a preference for situations in which the probabilities are known as opposed to those in which they are ambiguous in this sense, and this preference sometimes leads to behavior that appears to violate certain widely recognized axioms of rational choice (Ellsberg, 1961).

In what follows I am using ambiguity in its familiar connotation of having more than one interpretation or, more loosely, being obscure. Some statistical assertions that are ambiguous in this sense are ambiguous by design, their ambiguity is intended to serve some purpose; others are unintentionally ambiguous.

To illustrate a possible intentional use of ambiguity: If I wish to convince you of the durability of a particular foreign-made car, I might tell you that 90% of all such cars sold in the United States in the last 10 years are still on the highway. In fact, this claim says very little to the individual who thinks about it; in particular it does not tell one how long these cars have survived, even on the average. The assertion does not rule out the possibility that 90% of the cars of this make sold in the United States in the past 10 years were sold during the past year. The ambiguity can serve the purposes of the promoter of this make of vehicle very well however, if the average listener accepts the claim uncritically and, even better, misinterprets it to mean that 90% of all these automobiles are on the highway 10 years after they were purchased.

I received in the mail recently an advertisement from a major fertilizer distributor for a lawn treatment package with that distributor's product. Across the cover page in large print was the question: "Did you know that 92% of what liquid lawn care services apply to your lawn is water?" This is a classic example of the use of statistical innuendo. Let us assume that the claim is true; I have no reason to doubt that it is. It tells us nothing about the effectiveness of liquid lawn care services or whether they are better or worse than alternative approaches of comparable cost. It is quite possible that, given the chemicals that are used, 92% water is precisely the right mix. (I am not, of course, claiming that it is the right mix, but only that it could be.) Presumably the designer of this advertisement intended that the reader draw the conclusion that a liquid lawn service is a poor value relative to the alternative the advertisement is promoting, but the claim that it is 92% water provides no objective support for that conclusion. The lawn care advertisement is representative or many advertisements that use numbers in suggestive but completely uninformative ways.

Incidence statistics can be presented in various ways, depending on the perception one wishes to create. Sprent (1988) notes, for example, that in 1985 the number of road deaths per 100,000 population was about 9.1 in the United Kingdom and that the corresponding number for the United States was larger by about a factor of 2. On the other hand, when road deaths are reported per vehicle mile, the U.S. number is smaller than that of the U.K. Promoters of travel packages on opposite sides of the Atlantic would have different preferences as to which way this comparison is made.

If one wished to support the idea that coal mining had become safer during the period 1950 to 1970, one might point out that the number of accidental deaths per ton of coal decreased over that time. If one wished to make the case that coal mining had become a riskier vocation, one might note that the number of accidental deaths per employee increased. Both claims are true. Because mining became increasingly mechanized, the number of miners it took to produce a fixed amount of coal decreased over that period so it is possible for the number of accidental deaths per unit of product to go down while the number per working miner goes up (Crouch & R. Wilson, 1982).

Even when one uses the same set of numbers as the basis for reporting percentages or ratios, one may still be able to create quite different impressions depending on how one elects to report them. Dawes (1988) makes this point with reference to the reporting of the effects of smoking on health. One comparison involved 44.8 deaths per thousand among smokers and 21.1 per thousand among nonsmokers. If one wants to emphasize the detrimental effect of smoking, one is likely to report the death rate to be over twice as high among smokers as it is among nonsmokers: 44.8/21.1 = 2.12. However, if one wants to play down the effect of smoking, one might compare *survival* rates and point out that smokers are almost 98% as likely to live as are nonsmokers: 955.2/978.9 = .976. Many situations provide opportunities for reporting proportions or percentages in more than one way, thereby making it possible to create quite different impressions, depending on the representation one chooses.

Consider a case that led Dewdney (1993) to title a book on innumeracy and "math abuse" *200% of Nothing*. An ad for high-tech energy-efficient light bulbs and fixtures promised a 200% savings on energy resulting from the substitution of flourescent bulbs requiring 35 watts for incandescent ones requiring 100 watts. Most readers would probably agree that a reduction from 100 to 35 should be considered a savings of 65/100, or 65%, which is to say that if one wants to express a savings as a percentage, one should express it as a percentage of what the amount was before the reduction. The writer of the ad apparently chose to express the savings as a percentage of what the amount became after the reduction: 65/35, or roughly 200%. The latter sounds much more impressive, to be sure, but, as Dewdney points out, given the (presumably) conventional connotation of percentage saved, this would lead most readers to conclude that the effect of burning the fluorescent bulb would be not only to save all the energy consumed by the incandescent bulb but to generate as much again. This particular approach to making percentages seem more impressive than they really are Dewdney refers to as "percentage pumping."

We should note, however, that percentage calculations can be confusing for reasons other than the deviousness of advertising techniques. If I had $100 and gained or lost $10, we would say, I think, that I had gained or lost 10% of my original wealth. In fact, as already noted, the loss of a given magnitude may be perceived as larger than a gain of the same amount, because the former equals a larger fraction of the remaining wealth than does the latter. This sort of consideration led Daniel Bernoulli (1738) to distinguish between *fortune physique* and *fortune morale,* the former being the objective value of one's fortune and the latter its subjective worth.

A common formulation of the problem of solid waste in the United States is that 50% of all landfills now in use will close down within 5 years. As it happens, the same observation could have been made 20 or 30 years ago because most landfills are designed to be used for only about 10 years (Rathje, 1989). From the fact that half of those currently in use will be closed within 5 years nothing follows regarding how many areas will be in use at the end of that time. The number could be smaller than it currently is or it could be much larger. What may appear to follow, in the view of a reader who does not think deeply about the number, is that landfill areas are becoming scarce and that 5 years hence there will be only half as many as there are now.

One often sees in the reporting of sports news, items of the general form "such and such a team has won 8 of its last 10 games." Or "so and so has made the finals in three of his last four attempts." One can be reasonably sure when encountering such claims that the count started with a success. The truth of the assertion that such and such a team won 8 of its last 10 games does not rule out the possibility that the assertion that the team won 8 of its last 20 games is also true. The habit of reporting small-sample statistics in this way gives a somewhat misleading picture. A run of 8 wins in 10 tries may or may not be impressive, depending on how the sample was selected. Certainly it would be much more impressive if it were selected randomly than if one decided to start the count with a win, or worse, with a run of wins.

Often the reporting of the accuracy of diagnostic tests is ambiguous: Consider, for example, the claim that a particular test is 90% accurate. Does this mean: (a) given the disease, the test shows positive with a probability of .9, (b) given a positive test result, the disease is present with a probability of .9, (c) 90 % of all the test results, positive and negative, are accurate, or (d) something else. The differences among these possibilities can be substantial. Moreover, even if one knows which of them is intended, the situation may still be unclear. Tables 7.3 and 7.4 show two quite different situations, both of which are consistent with a claim of 90% accuracy in the first sense mentioned previously. Similar examples could be constructed relating to each of the other senses. For Tables 7.3 and 7.4 it is assumed that the disease has an incidence of 1 in 10,000 in the population. In the first example, the probability that the test shows (falsely) positive when the disease is absent is set at .01; in the second, it is set at .001.

In the first of these hypothetical examples, the probability that the disease is present, given a positive test result on a randomly selected person, is

$$p(D \mid P) = p(D\&P)/p(P) = .00009/.01 = .009,$$

TABLE 7.3

Results Expected With a Diagnostic Test That Correctly Detects 90% of the Cases of Disease and Erroneously Shows Positive on 1% of the Cases of No Disease

Test Result	*Ground Truth*		
	Disease (D)	*No Disease (~D)*	
Positive (P)	.00009	.01000	.01009
Negative (~P)	.00001	.98990	.98991
	.00010	.99990	1.00000

Note. Assumed incidence is 1 in 10,000. Cell entries are joint probabilities, assuming the test is administered to a large random sample of the population.

TABLE 7.4

Results Expected With a Diagnostic Test That Correctly Detects 90% of the Cases of Disease and Erroneously Shows Positive on .001% of the Cases of No Disease

Test Result	*Ground Truth*		
	Disease (D)	*No Disease (~D)*	
Positive (P)	.00009	.00100	.00109
Negative (~P)	.00001	.99899	.99891
	.00010	.99990	1.00000

Note. Assumed incidence is 1 in 10,000. Cell entries are joint probabilities, assuming the test is administered to a large random sample of the population.

about 9 chances in 1,000, or approximately 1 in 100. In the second case, the probability that the disease is present, given a positive test result on a randomly selected person, is

$$.00009/.001 = .09,$$

or almost 1 in 10. In both cases, remember that the test is 90% accurate in the sense that, given the presence of the disease, it returns a positive result 90% of the time. However, a random person getting a positive result would have only about one chance in a hundred of having the disease in the first case and about 1 chance in 10 in the second. (Of course people who get tested for a particular disease usually are not randomly selected from the population, but get the test because they have some reason to suspect they might have the disease, and these numbers would not apply to such nonrandom samples.) These examples illustrate the importance of false-positive rates, especially given low-incidence conditions, in interpreting the outcomes of diagnostic tests. They also serve as a reminder that ambiguities in the way test accuracy is reported lend themselves to erroneous conclusions of various sorts.

Statistical results can also be ambiguous, or meaningless, because of uncertainty about the way in which the samples on which they are based were selected. If there is a possibility that a statistical claim is based on a sample that was selected in a biased way, the claim cannot be taken to be representative of the general population to which the sample belongs. The results of polls in which the respondents are self-selected are highly suspect, for example, because people who voluntarily respond to a poll on a specific issue cannot be assumed to be a random sample from the population of potential respondents and therefore representative of the population as a whole with respect to the issue involved.

The general moral of this story is that statistical assertions are very often ambiguous, and sometimes even meaningless, and their ambiguity or lack of meaning may go undetected if they are interpreted uncritically without an effort to understand alternative interpretations they could have. Promoters know rather well how to exploit ambiguous statistical claims. The kinds of ambiguous claims that are sometimes made with the intent to mislead can also be made in good faith and an unawareness of their ambiguity. The challenge to the reader or listener to detect the ambiguity is the same in both cases, although one may feel rather differently toward the originator of an ambiguous statement that was intended to deceive than one does toward an individual who speaks ambiguously from statistical naiveté.

In a very readable little book that does not require a grounding in mathematics to be appreciated, Huff (1954/1973) has described numerous ways in which statistics—in the sense of the presentation of facts and figures—can be used, intentionally or unintentionally, to mislead. More recently, Gilovich (1991) has done so as well in a book with the delightfully ambiguous title *How We Know What Isn't So*. Other writers who have provided useful and highly accessible discussions of how statistics can be misused or misunderstood include Jaffe and Spirer (1987), Paulos (1990), and Dewdney (1993).

STATISTICS AS A THEORY OF MIND

> The idea of sampling is an essential element for making sensible decisions; indeed, it may be the basis of thought itself. (J. Cohen, 1957, p. 128)

Statistics appeared on the psychological scene in a major way first, following the work of Fisher and of Neyman and Pearson, as a tool for making inferences regarding the meaning of the outcomes of experiments. Perhaps because this application of statistics was seen to be very successful, psychologists began to use statistical inference procedures as metaphors for how people make decisions in daily life. Gigerenzer and Murray point to the application of the statistical theory of signal detection to human sensory and perceptual processes (Green & Swets, 1966; Swets, Tanner, & Birdsall, 1961; Tanner & Swets, 1954) as an especially noteworthy example of the fruitful use of this metaphor.

Gigerenzer and his colleagues (Gigerenzer & Murray, 1987; Gigerenzer et al., 1989) have suggested that psychologists have a tendency to make metaphors of mind based on tools and instruments that they find useful. Especially notable examples of such "tools-to-theories" transactions have involved statistics and computers. These metaphors, both of which emerged around 1960, have had a major impact on the subsequent history of psychological research. Not only did they provide conceptual models for perceptual and cognitive phenomena and frames of reference within which the phenomena can be viewed, they helped determine the kinds of research questions that were asked.

According to Gigerenzer and Murray (1987), as a consequence of the application of signal detection theory to the study of sensation and perception, and related work:

> [The mind, at least in certain functional contexts,] was now pictured as a statistician of the Neyman and Pearson school. The processes of "inference," "decision," and "hypothesis testing" were freed from their conscious connections and seen as unconscious mechanisms of the brain. Thus, uncertainty, in the sense of uncertain inferences and decisions, became an essential feature of cognitive processes, and computation of distributions and likelihoods, random sampling and power analysis became the mind's way of coping with this uncertainty. (p. 60)
>
> Problems were constructed so that they could be answered by *calculating* probabilities, means, variances, or correlations.... The new vocabulary for understanding human reasoning was the vocabulary of the statistician; the new elements of thinking were numbers (probabilities), and the process of thinking itself was explained by statistical operations such as calculating likelihood ratios. The theoretical questions asked by experimenters, the problems posed to the subjects, and the explanations sought all reflected the fascination with probability and statistics. (p. 147)

Although the mind-as-statistician metaphor was applied initially in psychology primarily to the study of sensory and perceptual processes, its domain of applicability was soon extended to include more cognitive processes as well. In recent decades it has prompted much effort to determine the abilities and limitations, the strengths and weaknesses, of people when faced with problems that require reasoning under uncertainty. The performance of participants in experiments is compared with ideal or normative performance, as defined by a specific theory of how such reasoning should be done. As has already been noted, researchers or interpreters of this research have not always recognized, or at least acknowledged explicitly, that the normative models against which human performance is judged are often themselves matters of dispute among statisticians and probability theorists.

When theoretical models have been used as normative standards against which to compare human performance, performance usually has been shown to deviate from the norms—to be suboptimal—in certain ways. Much of the effort of researchers in this area has been devoted to determining or demonstrating precisely how people's reasoning under uncertainty commonly deviates from what the normative models prescribe and explaining why it does so.

SUMMARY

There can be little doubt of the power of statistical methods for making inferences when they are thoughtfully applied with a clear understanding of their limitations and the assumptions that underlie them. But the cautions that have been raised by various writers (Abelson, 1995; Gigerenzer et al., 1989; Good, 1983a) about the dangers of applying these methods in cookbook fashion and the temptation to gloss over the complexity of the ideas in the interest of simplifying instruction or assuring the noncontroversial nature of a science's predominant methodology cannot be ignored.

Psychology, and other disciplines that make equally heavy use of statistical inference techniques, would be better served if the training of students seeking advanced degrees in the field put much more emphasis on the logic and frequently debatable assumptions that underlie the statistical procedures that have been developed for inferencing and less on the mechanics of the procedures themselves. Such training should include a thorough acquaintance with the history of probability theory and the philosophical questions that center on the concept of chance. Among the accounts of various aspects of this history and of many of the relevant issues are those of David (1962), Hacking (1965, 1975, 1990), Porter (1986), Gigerenzer and Murray (1987), Krüger, Daston, and Heidelberger (1987), Krüger, Gigerenzer, and M. S. Morgan (1987), Gigerenzer et al. (1989), and Salsburg (2001).

Training in statistics should also provide a perspective in which its use in hypothesis testing and decision making is viewed as one among other avenues to discovery. Macdonald (1997b) has characterized psychology as "a mix of evidence and theory held together by arguments" (p. 344). This seems right to me. The results of statistical tests are among the evidences from which arguments can be constructed, and they can be persuasive when used appropriately, but they are by no means the only ones, and when used inappropriately can be worse than no evidence at all.

In short, people who use statistics in their work to describe data sets or to make inferences need to understand their tools, the rationales on which they are based, and the assumptions that justify their use in particular instances. But it is not only people who use statistics in their work who benefit from knowledge of the subject. One must have some acquaintance with basic statistics if one is to read the newspaper comprehendingly, if one is to avoid being taken in by scams that capitalize on statistical naiveté, and if one is be an effective decision maker in an uncertain world.

CHAPTER

8

Estimation and Prediction

We cannot regard an action as rational unless it computes the probabilities.

—Hacking (1987b, p. 52)

In the preceding chapters, much attention has been given to a variety of topics relating to probability theory and statistics and relatively little to the results of research on how, and how well, people think about probabilistic matters. The results of some research have been discussed, but empirical studies have not been the major focus. Beginning with this chapter, attention shifts to research, its results, and the interpretations those results have been given.

Probabilistic or statistical reasoning has been studied in a variety of contexts including clinical diagnosis (Meehl, 1954), management decision making (R. V. Brown, Kahr, & C. R. Peterson, 1974), flood probability estimation by flood plain residents (Slovic, Kunreuther, & White, 1974), experimental design (Brewer & Owen, 1973), weather forecasting (Murphy & Winkler, 1974, 1977), climate change (National Academy of Sciences, 1983), accident analysis (C. H. Green & R. A. Brown, 1978; Slovic, Fischhoff, & Lichtenstein, 1978), election outcome predictions (Black, 1958; I. Fischer & Budescu, 1994), among many others. The review of this research here is not intended to be exhaustive, but it is extensive and the studies considered are intended to be broadly representative of the work in the field.

ESTIMATING SAMPLE STATISTICS AND RELATIVE FREQUENCIES

> If people were more capable of estimation and simple calculation, many obvious inferences would be drawn (or not), and fewer ridiculous notions would be entertained. (Paulos, 1990, p. 17)

The ability to estimate has not received the attention it deserves in education. It is an extraordinarily useful ability. Evidence that this is beginning to be realized is perhaps seen in one of six major recommendations in the 1980 agenda-setting report of the National Council of Teachers of Mathematics: "Teachers should incorporate estimation activities into all areas of the program on a regular and sustaining basis, in particular encouraging the use of estimating skills to pose and select alternatives and to assess what a reasonable answer might be" (p. 7). The importance of estimation skills has also been stressed by the Curriculum Framework Task Force of the Mathematical Sciences Education Board (1988). Although estimates can be made of many types of variables—for example, quantities, magnitudes, durations—here attention is limited to probabilistic or statistical variables.

Estimates of Central Tendency and Variability

When asked to observe a set of numbers and to estimate some measure of central tendency, such as its mean, people are able under some conditions to produce reasonably accurate estimates (Beach & Swensson, 1966; Edwards, 1967; C. R. Peterson & Beach, 1967), although systematic deviations from actual values have also been reported (N. H. Anderson, 1964; I. P. Levin, 1974, 1975). When the numbers whose means are to be estimated have been presented sequentially, effects both of primacy (influence of the first few numbers in the sequence) (Hendrick & Costantini, 1970) and of recency (influence of the last few numbers) (N. H. Anderson, 1964) have been obtained. For skewed distributions, estimates of means are likely to be biased in the direction of medians (C. R. Peterson & Beach, 1967).

People's ability to estimate variability has been studied but less extensively than their ability to estimate means. One focus of interest has been how perceived variability depends on the mean around which the variability occurs. Some investigators have reported results suggesting that perception of a distribution's variability is not influenced by the distribution's mean (I. P. Levin, 1975; Pitz, Leung, Hamilos, & Terpening, 1976). Others have found the two estimates to be related systematically. Estimates of the variability of a set of numbers have been noted to decrease as the mean of the set increased and the (absolute) variability remained the same; in other words, variances of the same magnitude around a small mean and around a large mean appeared larger in the former case (Beach & Scopp, 1968; Hofstatter, 1939; Lathrop, 1967). There is some question as to the extent to which this relationship reflects a true misperception as opposed to a confusion of variability in absolute terms with variability relative to a mean. A standard deviation of 20 pounds in the distribution of weights of 100 freight cars seems a good bit smaller than a standard deviation of 20 pounds in the distribution of weights of 100 people, even though in absolute terms it is not. It could also stem, at least in part, from the linguistic convention of making the interpretation of such words as *small* and *large* contingent on the context in which they are used, so what may be happening is analogous to when we consider a particular horse to be small or a particular dog large, even though in absolute terms the former is larger than the latter.

Given a variable set of values, one's perception of its variability may depend somewhat on the manner in which one becomes aware of the set. The variance is likely to be perceived as larger, for example, if the values are presented in random order than if they are presented in a regular ascending or descending order (Lathrop, 1967). More information on work that has been done on the estimation of sample statistics and on proposed models of the process by which estimates are made can be found in C. R. Peterson and Beach (1967), Pollard (1984), and Busemeyer (1990).

Estimating Relative Frequencies of Events

Relative frequencies of events often are perceived relatively accurately (Attneave, 1953; Carroll, 1971; Jonides & Jones, 1992; C. R. Peterson & Beach, 1967; Vlek, 1970; Zacks, Hasher, & Sanft, 1982), though not always (Fisk & Schneider, 1984). Several aspects of people's memory for frequency of occurrence, including the fact that it appears to be influenced little if at all by intention (Flexer & Bower, 1975; Hasher & Chromiak, 1977; Howell, 1973) led Hasher and Zacks (1979) to suggest that frequency information is stored in memory automatically. The same investigators have reviewed numerous studies that appear to support this view (Hasher & Zacks, 1984). In addition to its independence of intention, other aspects of memory for frequency that they cite as confirmatory of its automaticity include the lack of effects of training or feedback (Hasher & Chroniak, 1977; Zacks et al., 1982); the lack of sizable individual differences due to such factors as motivation and intelligence (D. Goldstein, Hasher, & Stein, 1983; Lund, Hall, K. P. Wilson, & Humphreys, 1983; Zacks et al., 1982), and the relative invariance of frequency memory with age (D. Goldstein et al., 1983; Hasher & Chromiak, 1977; Hasher & Zacks, 1979).

Several factors have been identified, however, that appear to be able to bias estimates in one or another way. Some of these have to do with how estimates are obtained. Estimates people produce of the frequency of occurrence of specific events in their own experience, for example, appear to be influenced by the way questions are framed. In one experiment, participants who were asked how many headaches they experienced per week gave larger estimates when the response alternatives provided for them were expressed as 1–5, 5–10, 10–15, ... than when they were expressed as 1–3, 3–5, 5–7, ... (E. F. Loftus, 1979). This result seems very strange indeed. How could the response scale convey, or even appear to convey, any information to a person regarding the number of headaches he or she had experienced? One possibility is that the scale that is provided is interpreted by people as the range of frequencies reported by the general population and is used to locate themselves on a continuum depending on whether they consider their own headache history to be extreme in either direction or somewhere in the middle of that range.

There is some evidence that estimates of the frequency with which some event (e.g., showing of a picture) has recently occurred can be increased (not necessarily made more accurate) by instructing people to imagine the occurrence of the event several times (M. K. Johnson, Raye, Wang, & J. H. Taylor, 1979; M. K. Johnson, T. H. Taylor, & Raye, 1977). This result seems similar to the finding that the act of imagining oneself experiencing a particular event can increase one's tendency to believe that one will actually experience that event (Gregory, Cialdini, & Carpenter, 1982). Forcing one to focus one's attention on an event appears to have the effect of increasing one's tendency to remember (or imagine) the event occurring in the past or to expect it to occur in the future.

Some work on frequency perception has focused on questions of how frequency is encoded and stored in memory and, in particular, on the level of organization at which storage occurs (Jacoby, 1972; Kellogg, 1981)—whether, for example, the frequency of occurrence of letters and words presented in an experiment is stored at the level of the individual letters (from which word frequency might be inferred) or at the level of words (from which letter frequency might be inferred) or both (S. J. Hoch, Malcus, & Hasher, 1986). Another focus has been on the role that attention plays in frequency perception and retention—is it necessary for an event to be attended to in order for its frequency of occurrence to be perceived and retained (Hasher & Zacks, 1979; Hintzman, 1986; Zacks, Hasher, & H. S. Hoch, 1986). One obstacle to getting a clear answer to this question, despite considerable experimentation, is the difficulty of defining precisely what constitutes a perceptual event (Johnson, Peterson, Yap, & Rose, 1989). For present purposes, it suffices to note that the results of research generally support the conclusion that people are quite sensitive to frequency of occurrence and can accurately distinguish between different frequencies under a variety of conditions.

Estimating Time and Costs

The problem of estimating how long it will take and how much it will cost to complete some task is a very common one in our society. Contractors must prepare time and cost estimates when bidding on projects. Local, state, and national government agencies must make similar estimates when allocating public resources to programs competing for the same funds. All of us, as individuals, estimate, for our own planning purposes, how long it will take us to do specific things and, in some cases, what the costs will be.

How good are we at making such estimates? The answer to this question has important implications for our individual and corporate lives. Especially interesting and potentially significant is the possibility of systematic biases in the way estimates are made and consistent errors in the results. If we are strongly inclined to underestimate, or to overestimate, the time and/or money that will be required to complete specific tasks, it should be helpful, for practical purposes, to know that. And if we do tend consistently to err in one direction, we would like to have an explanation of the fact.

A front-page article in the September 11, 1994, issue of the *Boston Globe,* entitled "The Big Dig," reports progress on the Third Harbor Tunnel project in Boston as of that time (Sennott & Palmer, 1994). Described in the article as "the largest public works project in America," the Big Dig involves 7.5 miles of construction including a tunnel providing access from south Boston under the Boston Harbor to Logan Airport. Estimates of the cost of the project grew, after the project was under way, from $360 million in 1976 to $7.74 billion in 1994. The latter estimate was the official one as of the date of the article, but the writers noted that, unofficially, officials were acknowledging that the ultimate costs were likely to be higher, perhaps as high as $10 or $12 billion. Official estimates of completion dates were 1998 as of February 1991, 2001 as of March 1992, and 2004 as of February 1993. A front-page article in the February 2, 2000, issue of the *Boston Globe* reported that Big Dig officials had just announced a $1.4 billion jump in the costs, bringing the then officially estimated total to $12.2 billion (Palmer, 2000); the same article gave early 2005 as the completion date, and described the project as being "roughly on track to be completed on time." The September 2001 summary report of the Oversight Coordination Commission of the Central Artery/third Harbor Tunnel Project (www.state.ma.us) gives $14.475 billion as the then current estimate of the total cost and a "substantial completion" date of 2004. (This report gives $2.3 billion as the initial cost estimate, made in 1984.) Some of the cost escalation is because the scope of the project increased very considerably over the years, some because of unanticipated complications encountered in the process of construction, and some because the reported figures do not correct for inflation, but even taking these factors into consideration, the history of this project is a graphic example of how very easy it is to underestimate by very large amounts cost and time-to-completion for complex projects.

The Third Harbor Tunnel project is far from unique with respect to its history of upward revisions of cost and time-to-completion estimates. Similar observations could be made about numerous major programs and projects. According to Gibbs (1994), time and cost estimates are chronically too low in major software development projects: "Studies have shown that for every six new large-scale software systems that are put into operation, two others are canceled. The average software development project overshoots its schedule by half; larger projects generally do worse" (p. 86). Pfleeger (1991) has made similar observations. One gets the impression from the news media that cost and time overruns (underestimates) are the rule in the government contracting world, if not in the world of project planning more generally; and, it would seem, the more complex the project, the greater the underestimation is likely to be.

But is this impression correct? It could be that the perception that overruns (and underestimates) are the rule rather than the exception is a misperception of the facts. Such a misperception could arise in a variety of ways. Perhaps overruns are more likely to be reported by the news media than are projects that finish on time and within budget. Even if this were not the case, it could be that reports of overruns, especially those that involve large amounts of money and some evidence or hint of scandal capture our attention and linger in our memories more than do reports of projects without these characteristics. Granted the possibility of the greater visibility and memorability of projects that exceed time and cost estimates than those that finish within time and on budget, I suspect that underestimation is a more common occurrence than overestimation.

I do not mean to single out contractors here as poor estimators of how long it will take to accomplish specific projects. My guess is that most of us are in this boat. I know from experience that when I estimate how long it will take to finish this book, I should make the best estimate of which I am capable, at least double it, and then not be surprised when, in fact, it takes considerably longer than that. My estimates of how long it will take to complete some personal chore seem almost invariably to miss the mark significantly on the low side; at least I am more keenly aware of misses of this type than of those of the opposite one.

Some research has been done on the question of how well people predict the time it will take them, or someone else, to perform a specified task. The results of this research support the notion that people generally underestimate how long it will take them to perform a specific task themselves, or they overestimate how much they can do in a given time (Buehler, Griffin, & McDonald, 1997; Buehler, Griffin, & M. Ross, 1994, 1995; Byram, 1997; Hayes-Roth & Hayes-Roth, 1979; Kidd, 1970), and this despite the fact that producing a specific estimate of how long it will take to complete a task may motivate one to complete the task within that time thus helping the estimate to become a self-fulfilling prophecy (Sherman, 1980). And underestimates sometimes persist despite efforts to improve them with debiasing techniques (Byram, 1997; Newby-Clark, M. Ross, Buehler, Koehler, & Griffin, 2000).

Interestingly, it appears that people are less likely to be overly optimistic when predicting the time it will take someone else to perform a task than when predicting how long it will take themselves to do so; they seem more likely to consider the possible effects of intrusions in the former case (Buehler et al., 1994; Newby-Clark et al., 2000). However, experts in the performance of a task may find it easy to underestimate how long it will take novices to do the same task (Hinds, 1999).

Assuming that there is a general tendency to underestimate, what might account for it? One possibility that should not be overlooked is that of the operation of motivational factors. Inasmuch as contractors typically bid for jobs and contracts are awarded to low bidders, any tendency they have to underestimate time and costs might be attributed, at least in part, to the incentive the bidding process provides to do so. Those who submit bids that take into account the high probability that not everything will go exactly as planned are likely to find themselves eliminated from the competition by others whose bids are based on much more optimistic, if less realistic, projections. Of course, if the bid is binding in the sense that the contractor must deliver what is promised for the estimated cost, this is a very risky strategy, but as it happens, in many contracts time overruns carry no penalty and cost overruns often are collectable, so the risk associated with underbidding may not be great.

Motivational factors could be involved in individuals' estimates as well as in those of corporate entities. Estimating a relatively long time to perform a task might be seen as an admission of lack of competence, for example, in which case one might be inclined to lower one's best estimate of how long a task will take, as opposed, say, to raising it. On this hypothesis, we underestimate how long it will take ourselves to do something because we tend to believe that we are more capable or more efficient than we really are, and we are more likely to underestimate how long it will take us than to underestimate how long it will take someone else, because we see ourselves as more capable and efficient than our peers. The latter idea is consistent with numerous other self-serving biases that research has revealed, such as the tendency to consider ourselves to have more positive and fewer negative character traits than our peers (Alicke, 1985), to be more fair-minded (Liebrand, Messick, & Wolters, 1986; Messick, Bloom, Boldizar, & Samuelson, 1985), less prejudiced (Fields & Schuman, 1976; O'Gorman & Garry, 1976), better equipped for academic success or marital happiness (Kunda, 1987), and better-than-average with respect to leadership ability (College Board, 1976–1977).

Another possible determinant of chronic underestimation is the difficulty we have in generating exhaustive lists of the members of vaguely defined sets (and/or overestimating the exhaustiveness of the lists we produce when we attempt to do so). In estimating the time required to complete a task, one is likely to imagine performing the task, which probably means going over it step-by-step in one's mind estimating the duration of each step. This process can go wrong and lead to an underestimate in either of two ways. First, one may overlook one or more steps. This possibility seems intuitively more likely than that of imagining extra steps that are not required. Second, in estimating the duration of any given step, one may overlook things that could go wrong or fail to think of complicating factors that could increase the time required to perform it. Again this possibility seems more likely than the offsetting one of factoring in to one's estimate complications and problems that do not arise. It appears that when people imagine how a future project may go, they typically do not imagine the various ways in which it might go wrong (Buehler et al., 1994). Providing people with an explicit list of what can go wrong may improve estimates in some cases (Engle & Lumpkin, 1992).

Evidence that people typically overestimate the probability of conjuctions of independent events and underestimate the probability of disjunctions (Bar-Hillel, 1973; L. J. Cohen, Chesnick, & Haran, 1971, 1972) also can help account for any general tendency to underestimate the time required to complete a task, especially a complex one. If a task is to be completed in minimum time, all of the several subtasks must go right (estimating the probability of their doing so is estimating the probability of a conjunction of events) and it must be that almost none of the possibilities for something to go wrong materializes (estimating the probability of one or more of these events happening is estimating the probability of a disjunction). If one is motivated to minimize estimates of time and costs, one is likely to do so for each of the subtasks. What is easy to overlook is that even if the probability of a delay-causing problem is small in each case, the probability that such a problem will occur in *at least one* of the components can be large. The assumption that everything will proceed as planned and that no contingencies will arise has been called by Kahneman and Tversky (1979b) the *planning fallacy.*

Another conjecture relating to task-time estimation is the following: When working on a nontrivial task that is spread over a significant period (say months), if one is asked at various times to estimate what percentage of the task has been completed, one is more likely to overestimate the percentage completed and to underestimate what remains to be done than the reverse. This conjecture rests on the assumption that we are likely to be more keenly aware of the details of the finished aspects of a task than of those of the remaining ones; we can call the former to mind because we have recently experienced them, but the latter are not so readily identified because we have to rely on our imagination to identify them.

A similar explanation can be given to the finding that people tend to remember better their own contributions to a group effort than the contributions of other members of the group, and therefore are likely to exaggerate the relative size of their own contributions (Brenner, 1973, 1976; Johnston, 1967; M. Ross & Sicoly, 1979). Such a bias may be accounted for by a form of the availability principle. One is bound to be more aware of the details of one's own work on a collaborative task, especially with respect to those aspects of the effort that are covert, than of the details of the work of one's collaborator(s). In recalling a project, one is likely to have available more information pertaining to one's own efforts than pertaining to the efforts of others, other things being equal.

Expressing Degree of Uncertainty

The problem of assessing people's intuitions regarding statistical or probabilistic variables is complicated by the fact that the terms that are commonly used to express degrees of uncertainty—"likely," "probable," "credible,"—and their complements, sometimes with qualifiers—"somewhat," "very," "extremely"—are not necessarily given the same connotations by all users, or even by the same user in different contexts (J. Cohen, Dearnaley, & Hansel, 1958; E. M. Johnson, 1973). So it is not always clear how such terms should be associated with probabilities, odds, or other concepts with precise quantitative connotations.

How the qualitative terms that people spontaneously use to express judgments of uncertainty or probability relate to quantitative expressions of probabilities has been the subject of some research (Beyth-Marom, 1982; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). There is some evidence that people may assess situations differently if asked to use numerical estimates of probability than if asked to express their assessments in qualitative terms (Teigen, 1988; Windschitl & Wells, 1996).

Experimenters have often, though not always, attempted to deal with the problem of terminology by explicitly giving the qualitative terms that are to be used in specific studies numerical meanings, but it is not always clear that people have followed the prescriptions. Some have asked people to express judgments directly in terms of probabilities or odds, but for many people these concepts are not familiar and it is not safe to assume that they are always used as intended.

People who get probability estimates from experts for use in risk assessment or policy analysis have developed approaches to probability elicitation intended to enhance the accuracy of the estimates obtained (M. G. Morgan & Henrion, 1990). These approaches have limited effectiveness, however, and it is not unusual for experts in the same field to produce quite different probability estimates especially when the probabilities being estimated cannot be checked objectively. Various methods have been proposed for aggregating the estimates of experts in order to derive a single best composite estimate for use in decision making or policy setting. Typically these methods involve averaging the estimates, possibly after weighting the individual estimates to reflect different degrees of expertise of the individuals who have produced them.

PREDICTING OUTCOMES OF PROBABILISTIC EVENTS

Some people make predictions, either explicitly or implicitly, in the performance of their jobs. Weather forecasters come immediately to mind as examples of people who are obliged to make explicit predictions on a daily basis. Stock brokers and investment advisers make predictions, at least implicitly, in advising investors regarding what and when to buy and sell. So do physicians in deciding which of several possible courses of treatment has the best prospects of success. Contractors must estimate the time and resources that will be needed to perform jobs on which they wish to bid; arriving at such estimates involves, at least implicitly, taking into account the probabilities of various contingencies arising. How good are the experts at predicting probabilistic events? Are they better than the rest of us? Or do they suffer the same difficulties and biases, and to the same degree?

Weather forecasters as a whole appear to be quite good at estimating the probability of rain, which is to say that it rains on about X% of the days for which forecasters predict an X% chance of rain (Murphy & Katz, 1983). The relatively high degree of accuracy in this case may be due to forecasters working within a system that rewards them for candor and provides them with constant feedback regarding the accuracy of their previous estimates (Henrion & Fischoff, 1986). How good experts in other areas are on predicting probabilistic outcomes is not so clear.

The Gambler's Fallacy

Possibly the best known misconception that relates to the ability of people to predict probabilistic events is the gambler's fallacy. This concept was discussed in chapter 2. It suffices here to recall that it is the belief that a run of successive occurrences of one type of random event (e.g., a run of heads in coin tossing) will make an additional occurrence of that event less likely, or that when a sample of ongoing random events shows a "deficit" of one type of event, the probability of the imminent occurrence of that event is increased. This fallacy has been known for a long time and has been demonstrated in many experiments.

Probability Matching Analogues

An experimental situation that has been widely used to study predictive behavior involves giving a person the task of predicting, on each trial of the experiment, the outcome of a probabilistic event, such as the color of the next flash of a light that has been flashing different colors, say red and green, on an irregular schedule. In this situation, people often predict a particular event, say "red," with about the same relative frequency as its actual occurrence and, as a consequence, end up with a smaller percentage of correct predictions than they could have obtained by simply predicting the more frequently occurring event on every trial (Estes, 1964; Kintsch, 1970; Meyers, 1976). If, for example, the light were flashing red on a random 70% of the trials, predicting red on 70% of the trials would result in a success rate of approximately (.7 × .7) + (.3 × .3) = .58, whereas predicting red on every trial would ensure a success rate of .7.

Analogous forms of suboptimal behavior have been observed in a variety of contexts in which people would do well to make all-or-none choices but use some other strategy instead. Even experienced poker players may make this mistake. Lopes (1976), for example, found that the sizes of the bets that players were willing to place increased in proportion to their subjective probability of winning. This is a suboptimal strategy. At least in the long run, the expected gain is maximized by betting the minimum possible amount whenever one believes one has a less than even chance of winning and the maximum possible amount whenever one believes one's chances of winning are better than even (N. H. Anderson, 1979).

Arkes, Dawes, and Christensen (1986) had people judge, on the basis of the number of A's that specific students had received in three courses, whether or not those students had graduated with honors. Participants were shown for each possible number of A's (0–3) the percentage of students receiving that number who graduated with honors. A minority of the students who received 0 or 1 A graduated with honors, whereas a majority of those who received 2 or 3 A's did. The percentages were such that participants who followed the rule of guessing honors for all students who had received 2 or 3 A's and not honors for all who had received 0 or 1 would have been sure of getting about 70% correct, and they were given this information. Many participants did not use this rule, however, and consequently got fewer than 70% correct.

This was especially true of those participants who had been explicitly encouraged to try to do better than 70% by being "extremely observant" and of those who had been motivated by the promise of a monetary reward for especially high performance. Those who were told that even experts could expect to do no better than 70% on the task and that people who tried to do better would actually do worse were more inclined to use the simple binary rule. Arkes et al. (1986) interpreted their results as demonstrating that an increased incentive to do well can cause a decrease in performance on a probabilistic judgment task. Such a finding is especially thought provoking in view of the fact that high motivation characterizes many of the circumstances under which professional diagnoses (say for medical or psychological purposes) are performed.

It will not escape notice that in this experiment use of the simple strategy of saying yes if the student has 2 or more A's makes the task trivially simple, and perhaps boring. One might decide to try to beat the odds (to do better than the 70% that is guaranteed by use of the two-or-more-A's-equals-yes rule), realizing that the probability of doing so is small. What one gains by this decision is participation in a real gamble with a long-shot chance of beating the odds; what one gives up is participation in a boring game. There is some evidence that even in casino gambling situations people may be motivated by the desire not only to win money, but to engage in interesting or exciting games (Keren & Wagenaar, 1985).

In a second experiment, Arkes et al. (1986) had people attempt to choose from among three candidates the baseball player who won the "most valuable player" award in the National League each year from 1940 to 1961 (excluding a few years in which pitchers won the award). Four items of information were provided for each candidate for each year: batting average, number of home runs, number of runs batted in, and standing of the player's team at the end of the year. All participants were informed that in about 75% of the cases choice of the player whose team finished highest in the standings would be correct.

Those who were only moderately familiar with baseball tended to make use of this fact and selected the player solely on the basis of his team's standing; those who were highly familiar with baseball did not follow this simple rule. The former participants did better than the latter as a group, although they expressed less confidence in their choices. Arkes et al. (1986) characterized the moderately knowledgeable participants as slightly underconfident and the highly knowledgeable ones as seriously overconfident, and noted that one of the dangers of overconfidence is the assumption that no assistance is needed to assure good decisions.

On the basis of a review of studies involving prediction in probabilistic situations, Meyers (1976) concluded that it is extremely difficult to get people to follow strictly an optimal strategy. One factor that may contribute to the failure of people to adopt optimal strategies in probabilistic prediction tasks is the positive reinforcement they receive on those trials on which their predictions prove to be correct. If it is true that people pay more attention to their successes than to their failures, then every time one makes a correct prediction, one's behavior, whatever it is, is reinforced. When one "probability matches" in probabilistic prediction tasks, one can get a fair amount of reinforcement and that may be enough to sustain the behavior. Indeed, as may be seen from Table 8.1 in the two-alternative situation, the difference between the guaranteed percentage correct with an optimal strategy and the expected percentage correct with a matching strategy is not enormous even at its maximum, when the probabilities of the two events are .75 and .25, and it is quite small when the probabilities of the two events are either very different or nearly the same.

This line of reasoning gets some support from the finding that item-by-item feedback on the outcomes of probabilistic choice tasks sometimes results in poorer performance than complete lack of feedback (Arkes et al., 1986; Hammond, Summers, & Deane, 1973; Schmitt, Coyle, & King, 1976). Feedback may encourage people to focus on individual trials and to make each prediction contingent on the outcome of the immediately preceding trial, using some simple strategy such as "stay with same prediction following successful trial, change prediction following an unsuccessful one." Such a strategy would, of course, result in probability matching. It also may be, as Dawes (1979) has suggested, that people feel compelled to try to account for all of the variance in such tasks even though it is not possible to do so.

It could also be that college students participating in probabilistic prediction experiments are unwilling to use optimal strategies, and clinicians resist using algorithmic approaches to diagnosis, for the simple reason that, in both cases

TABLE 8.1

Guaranteed and Expected Success Rates for Optimal and Matching Strategies, Respectively, in Two-Alternative Event Prediction Tasks

		Strategy	
Alternative Probs	*Optimal*	*Matching*	*Difference*
.55/.45	.55	.505	.045
.65/.35	.65	.545	.105
.75/.25	.75	.625	.125
.85/.15	.85	.745	.105
.95/.05	.95	.905	.045

decision by formula takes the challenge out of the task. Parenthetically, we should note that when the two events are equally probable, .5/.5, there is no optimal strategy; the expectation is for a .5 success rate no matter what sort of prediction strategy one uses. There is some evidence, however, that many people believe that one should be able to do better than chance when trying to predict, say, the outcomes of tosses of a coin—that performance should improve with practice and be impaired by distractions (Langer & J. Roth, 1975).

Predictions and Preferences

> Our passions, our prejudices, and dominating opinions, by exaggerating the probabilities which are favorable to them and by attenuating the contrary probabilities, are the abundant sources of dangerous illusions. (Laplace, 1814/1951, p. 160)

Suppose you are to enter a lottery in which the probability of winning is .1. Should you care whether the .1 represents there being one winning ticket in 10, 10 winning tickets in 100, 100 winning tickets in 1,000, ... ? As long as 1 ticket in every 10 sold is a winner (and wins the same amount), what should it matter how many tickets are sold? It appears that many people do have a preference for a larger number of tickets and winners, for a fixed probability of winning. Kirkpatrick and Epstein (1992) demonstrated this, for example, with an experiment in which people who were to win a prize if they drew a red bean from a bowl preferred to draw from a bowl that contained 100 beans 10 of which were red than from a bowl containing 10 beans 1 of which was red. Denes-Raj and Epstein (1994) found that some people even preferred to draw from a bowl that contained 100 items, from 5 to 9 of which were red, than from a bowl containing 10 items, 1 of which was red; these people actually preferred the situation that provided the lower probability of winning.

When a compound event is composed of two or more independent component events, the probability of the compound event is the product of the probabilities of the component events, and the order of occurrence is irrelevant. That order of event occurrence in such cases may not be seen as irrelevant was shown by an experiment by Rowen (1973) in which participants were required to choose one of two mutually exclusive actions both of which had the same probability of success. However, both actions were composed of two independent steps, the probability of success of which differed. Participants tended to select the action composed of steps the first of which had the greater probability of success.

In most laboratory studies of the ability of people to predict probabilistic events, the events are of no great consequence or intrinsic interest to the participants. We can probably safely assume that, for the most part, people have no strong preference for the occurrence of one of the possible events over that of the other. Presumably they want their predictions to be as accurate as possible, because making correct predictions is their task, but, apart from this consideration, whether the red light comes on more frequently than the green one is of little concern.

In many real-life situations involving probabilistic events, we care a lot about outcomes and strongly prefer some of the possibilities over others. We care, for example, about how likely we are to end up as a highway fatality statistic, about the probability that the global temperature really is on the rise, about the prospects of an improved economy, maybe even about the likelihood that the weather will be good tomorrow. Do our preferences for particular outcomes over others affect our ability to estimate the probabilities of uncertain events? And if so, how much of an effect do they have, and what is the nature of that effect?

Sundstrom, Lounsbury, DeVault, and Peele (1981) took opinion polls of people living in a small community near the planned site of one of the world's largest nuclear-power plants as of 1980. They related their results to an expectancy-value attitude model that assumed one's attitude toward some object, in this case the power plant, is a simple sum of the possible positive and negative consequences of the object's existence each multiplied by the individual's expectation of its occurrence. The results supported such a model inasmuch as expressed attitudes reflected people's opinions about the relative probabilities of positive and negative consequences of the plant's existence. People did not give equal weight to all possible consequences, however, but tended to overweight some and underweight others. The correlation between the degree to which one liked or disliked a potential consequence and its estimated probability of occurrence was high, suggesting that the two judgments were not independent. In a very different context, Babad and Katz (1992) showed that soccer fans were likely to overestimate the probability that their favorite team would win a specified game.

The idea that how desirable, or undesirable, one considers an event to be may affect one's estimate of the probability of its occurrence finds considerable support in the experimental literature (S. J. Hoch, 1985; Irwin, 1953; MacCrimmon, 1968; Marks, 1951;. Slovic, 1966; Weinstein, 1980). When asked to predict cards drawn at random from packs containing cards representing different payoffs, people often tend to overestimate the likelihood of drawing the more desirable cards (Irwin, 1944, 1953; Irwin & Snodgrass, 1966).

There is also some evidence that people tend to overestimate the probability of an event to which their attention has been directed—a focal event. Gibson, Sanbonmatsu, and Posavac (1997) found that people were more likely to bet on a national basketball team they had been instructed to imagine winning than to bet on any of a set of competing NBA teams. These investigators see this behavior as an example of the more general finding of selective hypothesis testing whereby people tend to overrate the likelihood of a hypothesis that they are considering relative to those of competing hypotheses that could be considered.

Lotteries

Laplace (1814/1951) argued that people who play at lotteries generally do not understand the odds against their winning: "They see only the possibility by a small stake of gaining a considerable sum, and the projects which their imagination brings forth, exaggerate to their eyes the probability of obtaining it; the poor man especially, excited by the desire of a better fate, risks at play his necessities by clinging to the most unfavorable combinations which promise him a great benefit" (p. 161). Laplace also pointed out that the player's exaggeration of the prospects of winning is bolstered by the publicity that winning receives and that losing does not: "All would be without doubt surprised by the immense number of stakes lost if they could know of them; but one takes care on the contrary to give to the winnings a great publicity, which becomes a new cause of excitement for this funereal play" (p. 161).

Do people who play lotteries really understand the odds against their winning? Probably not. Certainly promotions of lotteries, including those run by the state, are not designed to stress the fact that almost everyone who plays loses. Publicity that follows drawings invariably focuses on winners. Seldom, if ever, does one hear or read of the many people who, week after week, put money that they can ill afford to lose into the black hole of a public lottery in the unrealistic hope of striking it rich.

Would people play lotteries if they did understand the odds? Perhaps some who now do so would not, but undoubtedly many would. No matter what the odds, and how well they are understood, one can always take the position "Someone has to win, and my chances are as good as anyone else's."

Is this a rational position? If one defines rationality in terms of the maximization of expected value, it clearly is not. Lotteries are designed to generate income for the entities (state, organization) that run them, which means that the payout to the winner(s) must be less (and typically is considerably less) that what is taken in from participants. It follows that the expected value of a ticket (the product of the amount that would be won and the probability of winning) must be less (and it typically is considerably less) than its cost. So when one buys a lottery ticket for $1.00, one, in effect, pays $1.00 for something with an expected value of, perhaps, 50 cents. Not very rational behavior from an expected-value point of view.

But one might argue that this is a much too simple-minded assessment of the situation. It completely ignores the value of the pleasure one gets from playing the game. People spend money all the time on activities from which they have *no* expectation of a monetary return (movies, concerts, roller coaster rides, trips around a golf course). Like many activities, playing the lottery has entertainment value, win or lose.

Compulsive gambling is not a new phenomenon, but as state lotteries have become increasingly common, it appears to have become an increasingly noticeable and troublesome problem with societal ramifications. It is a bit ironic to hear enticing promotions of state-run lotteries and public-service announcements of state-sponsored clinics for compulsive gambling nearly juxtaposed on the airways. It would be good to know more about how people's tendency to be overconfident of their ability to beat the odds, especially when bolstered by lopsided reporting or lottery outcomes, contributes to this problem.

Sample-Size Intuitions

Researchers have focused on both the question of whether the law of large numbers—which refers in this context to the tendency of large random samples, but not small ones, to resemble the populations from which they were drawn—is reflected in the intuitions of people untrained in statistics and on that of the extent to which people, such as research psychologists, who have had such training make statistical decisions that are inconsistent with it. Evans (1989) points out that this research has concentrated on the isssue of whether people overestimate the power of small samples and that little, if any, attention has been given to the question of whether people understand the diminishing returns to be obtained from ever-increasing sample size. This is unfortunate, he suggests, because "optimal decision behavior requires one to sample sufficiently but not excessively" (p. 34).

Among what appear to be faulty intuitions that have been observed about probabilistic events is the expectation for small samples to resemble too precisely the populations from which they are drawn. For example, it has been claimed that when people are asked to predict a sequence of tosses of a fair coin, the proportion of heads in arbitrary short segments of the sequences they produce tends to be closer to .5 than would be expected according to probability theory (Tune, 1964). Tversky and Kahneman (1971) refer to the intuitions behind such judgments as the "law of small numbers," according to which small samples are assumed to be highly representative of the populations from which they are taken. (See also Wagenaar, 1970, 1972.)

During the first couple of weeks of a major-league baseball season, the leading batting average is likely to be between 400 and 500. By the end of the season it is almost certain to be under 400. An interpretation that is sometimes put on this fact is that batters get worse or pitchers improve as the year wears on. A more plausible account recognizes that small samples are more likely than large ones to produce deviant numbers. This explanation could be tested by computing batting averages over 2-week periods throughout the season. On the sample-size hypothesis, one would not expect the leading 2-week average to decrease systematically over the year. Of course there may be systematic trends superimposed on the sample-size effect, but the point is that even in the absence of such trends the sample-size effect would be expected to yield the observed decrease in the leading average, when that average is computed on a steadily increasing sample size.

Kahneman and Tversky (1972) have observed that people often do not take sufficient account of sample size when estimating probabilities for which sample size is a relevant consideration. The following problem illustrates the point:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know about 50 percent of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.
>
> For a period of one year, each hospital recorded the days on which (more/less) than 60 percent of the babies born were boys. Which hospital do you think recorded more such days? (p. 443)

Of those participants who selected one hospital or the other (they were also allowed to indicate indifference between them) about half selected the one with the larger number of births. The result appears to indicate a lack of understanding of the fact that large deviations from a mean, in percentage terms, are more likely in small samples than in larger ones.

Several subsequent studies have shown that people's performance on such problems can be quite sensitive to the way in which they are worded (Bar-Hillel, 1979; Evans & Dusoir, 1977; C. L. Olson, 1976), and this raises the possibility that what is sometimes attributed to faulty statistical intuitions could be due, at least in part, to a failure to comprehend the problem as given. Evans and Dusoir (1975) found, for example, that when participants were asked which of the two hospitals described previously was more likely to have a day on which all babies born were boys, and were forced to select one or the other, 85% selected the smaller hospital. When asked directly whether an accurate result was more likely to come from a statistical estimate based on a small sample or on a large one, nearly all parcipants correctly indicated the large sample.

Some research psychologists who have had formal training in statistics and experimental design believe that a sample randomly drawn from a population will be similar to the population in all important respects (Tversky & Kahneman, 1971). This leads to the expectation that two randomly drawn small samples from the same population will be more similar to the population and to each other than sampling theory would predict, and to such errors of statistical reasoning as overestimation of the replicability of results from a single experiment and failure to consider the possibility that unexpected results could be due to sampling variability. In general, it appears that people often are unaware of or ignore the importance of the validity of the information on which predictions are based and are unaware that in the case of low validity, predictions should be regressed toward some central tendency measure such as the mean (Kahneman & Tversky, 1973).

L. J. Cohen (1962) has reported evidence that the statistical power of many psychological experiments is extremely low, which means that by using small samples experimenters run a high risk of failing to detect effects that are real, which is to say, they risk failing to reject the null hypothesis when it is false. Tversky and Kahneman (1971) obtained evidence on this issue by asking professional psychologists a variety of questions regarding the design of studies intended to replicate or check previously obtained experimental results. In general, the sample sizes the respondents specified for replication studies were sufficiently small to guarantee a large percentage of failures to find effects that were there to be found. In Tversky and Kahneman's words, the believer in the law of small numbers "gambles his research hypothesis on small samples without realizing that the odds against him are unreasonably high" and in "evaluating replications, his or others, he has unreasonably high expectations about the replicability of significant results" (p. 109).

Somewhat in contrast to the evidence that indicates the existence of an intuitive law of small numbers, there is also evidence that people have some sensitivity to the importance of sample size in statistical reasoning and are able to apply this sensitivity in a discriminating way (C. R. Peterson, DuCharme, & Edwards, 1968). Expressed confidence in estimates of the mean or variance of a set of numbers tends to increase (confidence intervals decrease) as the sample size increases (Bar-Hillel, 1979; DuCharme & C. R. Peterson, 1969; C. R. Peterson & Beach, 1967). People seem to understand also that one needs a larger sample to justify a generalization about a characteristic that is likely to be highly variable in the population than to justify a generalization about a characteristic that is likely to be more constant (Evans & Pollard, 1985; Nisbett et al., 1983). (Understanding of this principle does not preclude stereotyping, which can rest on underestimation of within-group variability with respect to specific characteristics.) How likely people are to use the law of large numbers effectively in reasoning depends, in part, on whether the situation about which they are reasoning seems to be governed by chance effects to a substantial degree (Jepson, Krantz, & Nisbett, 1983). For example, people are less likely to believe that a small sample of a slot machine's behavior is representative of its long-term behavior than to believe that a small sample of an athlete's performance is representative of his or her general ability.

It appears also that people are more likely to use the law of large numbers when the events involved are highly "codable" than when they are not (Kunda & Nisbett, 1986a; Nisbett et al., 1983). To be codable in this context means roughly to be counted easily or to have features or aspects that are easily counted. Kunda and Nisbett use sports activities as examples of codable events. A basketball game has associated with it many events that can be counted, such as number of baskets, rebounds, or assists per player. Social behavior is not likely to be so easily coded. Kunda and Nisbett suggest also that people are more likely to use the law of large numbers for familiar domains and problem types than for unfamiliar ones, which may be because more familiar events tend to be more codable.

Population-Size Intuitions

By comparison with the research that has been done on intuitions regarding sample size, relatively few studies have addressed the question of intuitions about population size or, more specifically, about sample-to-population ratio. The few relevant studies that have been reported suggest that people tend to believe that the larger the population, the larger the sample that is needed to represent it adequately, or, conversely, that small samples are likely to be more appropriate for smaller populations than for larger ones (Bar-Hillel, 1979; Evans & Bradshaw, 1986); Evans and Bradshaw's participants selected population size as the most important of several variables as determinants of how many batteries should be sampled from a truckload of either 10,000 or 20,000 to decide whether the entire load satisfied a statistical acceptance criterion. Some readers may be surprised to learn that, within broad limits, the adequacy of a sample is independent of the size of the population from which it is drawn.

"Within limits" is an important qualification here, because it is not the case that population size is never relevant to the question of adequate sample size. A random sample of modest size that almost exhausts the population from which it is drawn is likely to represent its population somewhat better than a random sample of equal size drawn from a much larger population. In general, however, in situations in which random sampling makes practical sense, the sample will be a relatively small proportion of the population, and in these cases what constitutes a sample size that is adequate to provide an estimate of a parameter with a given level of expected accuracy is essentially independent of the size of the population. A random sample of 1000, for example, is likely to provide as accurate a representation of a population of 100,000 as of a population of 10,000.

L. J. Cohen (1982) makes a distinction between confidence and weight, and argues that confidence is determined strictly by sample size but that weight is determined by sample-to-population ratio. Confidence, according to Cohen's analysis should depend on the proportion of random samples that resemble (within specified limits) their parent populations with respect to some measure of interest, and this is strictly a function of sample size, at least for samples that are small relative to their populations. In contrast, weight or strength of evidence reflects, at least in part, whatever legitimate reasons one has to assume that the sample is not biased in any way, and one way to increase the probability that it is not is to increase the fraction of the total population it contains. Again, this is easiest to see in the case of samples that are relatively large fractions of their parent populations; there simply are fewer opportunities for bias with a sample that contains 90% of its population than with one that contains 10% of its.

RISK ASSESSMENT, COMMUNICATION, AND PERCEPTION

Risk assessment has to do with assigning probabilities to future events that people would like, if possible, to avoid; it refers to the work that professionals do in attempting to quantify specific risks. Risk communication involves various methods for conveying information about risks. Risk information is communicated sometimes by experts, but also sometimes by nonexpert members of the press. Risk perception is subjective and personal; the same risk (in terms of incidence statistics) may be perceived very differently by different individuals. A lack of correspondence between specific risks as assessed by professionals and as perceived by the public has been seen as the source of considerable trouble (Hance, Chess, & Sandman, 1988).

Some of the risks that concern people have to do with events—automobile accidents, heart attacks, strokes—that have occurred many times in the past. In these cases, actuarial data are useful in assessing risks. Some risks must be assessed independently of frequency data or estimates, because they relate to events that have never happened in the past, such as the risk of a global nuclear war or the risk of the world population becoming catastrophically large. Many risks also are difficult to quantify because of a lack of knowledge of a fundamental nature; the effects of long-term exposure to many (perhaps most) chemicals that are used in various industrial, agricultural, medical, and military contexts, for example, are not known (Shodell, 1985). In some instances, relevant events have occurred, but not sufficiently often to provide a stable statistical basis for predicting their likelihood in the future. In other cases, such as the controversial question of the risk of cancer from indoor radon, the relatively few studies that have been done have yielded conflicting results (Horgan, 1994).

Attempts to quantify risks for which frequency data do not exist, even when made by experts, can disagree by large amounts (Hammond & Marvin, 1981; Lowrance, 1976). Apparently experts disagree with respect to the relative importance of various factors that, in combination, determine the magnitude of a specific risk. There is some evidence that the interjudge variability can be reduced by the use of structured analytic judgment processes, but whether greater interjudge agreement means more accurate judgments in those cases in which there is no independent objective check on accuracy is not clear.

Risk assessments are often performed for the purpose of guiding the formulation of policy aimed at avoiding actions that are likely to create major problems for the future, or at identifying actions that could improve future outlooks from a risk-assessment perspective. Such assessments influence legislation, regulatory policies, strategic defense planning, research activities, investment decisions, and many other aspects of our corporate and individual lives. The history of efforts to assess and manage risk has been engagingly told by P. L. Bernstein (1996).

Few attempts at risk assessment in modern times have received more attention than those associated with the nuclear power industry. Following numerous studies, the Nuclear Regulatory Commission (NRC) established design objectives that specified acceptable risk from nuclear-power plant operation in quantitative terms. Specifically, it stated as an objective that the risk to an individual in the vicinity of a nuclear-power plant of being killed as a consequence of a reactor accident should not exceed one 10th of 1% of the sum of the risks of being killed by other accidents to which members of the U.S. population are exposed (Nuclear Regulatory Commission, 1986).

This is a strict standard of safety. In 1990, an estimated .037% of the U.S. population died of accidental causes, including no one, I believe, from a nuclear-power plant accident (M. Hoffman, 1991). Assuming this figure is relatively constant from year to year, the NRC standard, if met, means that the probability that an average citizen in the vicinity of a nuclear-power plant dying as the result of a power plant accident in a given year should not exceed .000037%, which corresponds to a probability of .00000037, or less than 1 chance in a million. If one 10th of the U.S. population lived "in the vicinity of a nuclear-power plant," which it does not, this would give an expectation of about 9 deaths per year, on average, from nuclear-power plant accidents.

To put this in perspective, one must remember that the annual death toll in the United States from motor vehicle accidents is approximately 40,000. According to Bartecchi, MacKenzie, and Schrier (1995), smoking-related illnesses accounted for more than 400,000 of the more than 2,000,000 deaths in the United States in 1990, and for more than one quarter of all deaths among people aged 35 to 64. Given such numbers, why, one might ask would anyone object to the building of nuclear-power plants; why are people who are concerned about risks to human life not putting their energies into protesting the manufacture of automobiles or cigarettes—demonstrated killers—rather than into the building of nuclear plants? The situation is, of course, more complicated than the numbers suggest—objections are based not only on concern about immediate effects of possible accidents but on long-range problems associated with storage and disposal of radioactive material. A major uncertainty that no one knows how to resolve to everyone's satisfaction is that of how to ensure that the NRC's probabilistic "one 10th of 1%" objective has been met.

Policy often must be formulated without the benefit of objective data on which to base estimates of probabilities because such data simply do not exist. There is no alternative in such cases to that of relying on the judgments of people who, by virtue of training or experience, seem to be in the best position to have opinions that should carry weight. Of course, the problem of deciding whose opinions should carry weight is itself a matter of judgment, as is the problem of deciding who should make *this* decision, and so on. But policy analysts do not concern themselves about the risks of infinite regresses and they manage to cope with uncertainties primarily through reflection, dialogue, and consensus building. As M. G. Morgan and Henrion (1990) point out, policy analysts and policymakers seldom have the luxury to be fastidiously scientific in all respects, the need for policy decisions is gated by events that occur on their own schedule and that do not wait for opinions to be verified by experimental results.

Risk assessment and perception have been the focus of considerable psychological research (Apostolakis, 1990; Fischhoff, Lichtenstein, Slovic, Derby, & Keeney, 1981; Fischhoff, Sverson, & Slovic, 1987; McCormick, 1981.) The results from several studies show that the perceived relative riskiness of various situations does not correspond closely to the actual riskiness of those situations, at least as reflected in incidence statistics (Slovic, Fischhoff, & Lichtenstein, 1979.) It also appears to be the case that the general public's perception often differs from that of experts with respect to specific risks (P. A. Bell, J. D. Fisher, Baum, & Greene, 1990; Burton et al., 1978; Kempton, 1991; Slovic, Flynn, & Layman, 1991; Weart, 1988).

The question of the reasonableness of the attitudes of the general public about risks for which frequency data do not exist is complicated by disagreement among experts as to what the risks really are. Consider, for example, the question of global warming. Some experts believe the possibility that the average temperature of the earth is being raised as a consequence of the accumulation of greenhouse gases in the atmosphere to be among the more serious long-term threats that humankind faces (Bolin & Doos, 1986; R. E. Dickinson & Cicerone, 1986; Houghton & Woodwell, 1989; Kerr, 2000). Others have taken the position that the evidence of a real threat is weak and that the risk has been greatly overstated (Seitz, Jastrow, & Nierenberg, 1989). Debate on the issue continues (Roberts, 1989; White, 1990), and policy decisions must be made long before it is likely to be resolved by the acquisition of enough data to make the long-term trend crystal clear.

Perception of the Relative Seriousness of Specific Risks

In one study of risk estimation, both physicians and college students overestimated the risk of death from various specified diseases, although the physicians' estimates were more nearly accurate than those of the students (J. J. J. Christensen-Szalanski, Beck, C. M. Christensen-Szalanski, & Keopsell, 1983). For both groups, estimates were directly related to the frequency with which the diseases had been encountered. The investigators concluded that experts and nonexperts probably use similar thought processes to make frequency estimates but that they differ with respect to their exposure to the estimated events.

A specified risk is likely to be perceived the more serious the larger the number of deaths or critical injuries that could result from a single incident (C. H. Green & R. A. Brown, 1978). Also the frequency of sensational causes of death (homicide, tornado) tends to be overestimated, whereas that of more mundane causes (asthma, diabetes) is typically underestimated (Lichtenstein, Slovic, Fischoff, Layman, & Coombs, 1978). These findings may be attributable in part to the greater memorability of incidents affecting many people at one time in attention-getting ways and of events with sensational causes, and they may rest in part also on the tendency of the news media to give greater coverage to these types of events (Combs & Slovic, 1979). Suggestive evidence that people's estimates of the relative frequency of various causes of death are influenced by the amount of media coverage given to them may be seen in the fact that accidental death is likely to be estimated to be more frequent than death from stroke, although the reverse is true in fact (Slovic, Fischoff, & Lichtenstein, 1976).

Reporting by the media of information regarding risks tends to focus on newsworthy events—often catastrophic incidents such as the 1986 nuclear-power plant meltdown in Chernobyl or the 1984 toxic-chemical release in Bhopal. It is not necessary to deny the importance of such events to observe that media focus on them can help foster inaccurate beliefs about the relative magnitudes of various threats to the environment. The perceived risk of nuclear contamination increased considerably, for example, following the Chernobyl incident (Midden & Verplanken, 1990; Renn, 1990; van der Pligt & Midden, 1990; Verplanken, 1989). More generally, the public's attitudes about various man-made health hazards—for example, possible carcinogens—appears to be highly correlated with their frequency of mention by the media, which often is at variance with the views of the scientific community (Rothman & Lichter, 1996).

It may be that rare events receive more publicity than frequent events in part because of their rarity, and that they consequently come to be seen as more commonplace than they are. Paulos (1990), who makes this point, notes that sensationalistic events—terrorist kidnappings and cyanide poisonings—are given a great deal of media coverage, often of a highly emotional sort, whereas more mundane tragedies, such as the 300,000 deaths per year from smoking in the United States alone—"roughly the equivalent of three fully loaded jumbo jets crashing each and every day of the year"—are given little attention by comparison.

The effect of media attention is also illustrated by the crash of the *Hindenburg* in Lakehurst, New Jersey, in 1937. This incident effectively brought to a halt the use of zeppelins as a means of transportation and interest in developing the technology it represented, despite the fact that the 36 people who lost their lives in this tragic event were the only people to die as the result of a mishap during 20 years of commercial airship travel (McPhee, 1973/1992). Only very recently, more than half a century after the disaster, is interest in zeppelin-type airships of transportation being revived (Hangenlocher, 1999).

The history of lighter-than-air aircraft stands in rather striking contrast to that of steamships. More than 3,000 people were killed as the result of explosions of steamship boilers between 1816 and 1851. (Burke [1966] gives an account of how the commonality of such explosions gradually changed the public's attitude, and that of their elected representatives, about governmental regulation of private enterprise.) One must assume that the fact that the *Hindenburg* incident was captured on film as it happened, and was widely publicized immediately following its occurrence and for years thereafter, had much to do with the association of disaster with this type of air travel in the public's mind and with the abrupt change of attitude regarding lighter-than-air aircraft as a means of transportation.

Extensivesness of media coverage undoubtedly is a major factor in shaping public opinion about the relative seriousness of various risks. It seems unlikely, however, to be the whole story. It does not account, for example, for the finding that experts commonly overestimate both the dangerousness of events and the frequency of occurrence of dangerous events in psychiatric contexts (Ennis & Litwack, 1974; Steadman & Cocozza, 1974; Ziskin, 1975). Sensational, emotion-evoking events are more likely than more bland events to capture our attention when they occur and our imagination when they do not, and to be remembered in either case.

Irrelevant Influences on Risk Estimations

Direct evidence that frequency estimates are sometimes influenced by factors that are independent both of the actual frequency of occurrence and of frequency of exposure to media reports comes from a study of mood effects by E. J. Johnson and Tversky (1983). Participants were given descriptions of individual deaths due to various diseases, natural disasters, accidents, and crimes. These descriptions were written as short newspaper stories, and contained no information about prevalence. Participants were then asked to estimate the number of annual fatalities due to specific causes. Data were also collected on the interest, quality of writing, and mood evoked by each story. The investigators were particularly interested in how risk assessment would be influenced by the stories' affective effects. They found a strong generalized effect of induced mood on frequency estimates (all perceived risks affected approximately equally) and no hint of local effects (effects unique to the specific risk identified in the story) or a generalization gradient (with risks being affected in proportion to their similarity to the specified risks). Johnson and Tversky replicated the effect for nonfatal risks and for positive affect, and found that perceived risks were generally reduced. They concluded that the mood induced by brief reports has a large and pervasive impact on estimates of the frequency of risks even when the risks being judged are unrelated to the cause of the mood.

Mood effects on judgment have been observed in contexts other than risk assessment, as conventionally understood (Bower, 1995). People tend, for example, to judge their own behavior, attitudes, and abilities more positively when in a good mood than when in a bad one (Forgas, Bower, & Krantz, 1984; Sedikides, 1992). Estimates of success on a completed task may vary with the mood one is in when the estimate is made, better moods being associated with higher estimates of success. Such findings leave open the question of cause and effect—a better mood could induce higher estimates, or a higher estimate could improve one's mood, or both one's mood and one's estimate could be, at least in part, results of one's actual performance on the task.

There is some evidence of a general tendency for people to underestimate their personal vulnerability to various types of risk, or at least to estimate their own vulnerability to be less than that of their peers (Weinstein, Klotz, & Sandman, 1988; Weinstein, Sandman, & Roberts, 1991). Or, to state this type of bias in positive terms, people have been shown to be optimistic about their own futures relative to their expectations regarding the futures of others (Weinstein, 1980, 1982, 1983, 1984, 1987, 1989; Weinstein & Lachendro, 1982). It appears that we tend to consider specified positive and negative events to be, respectively, more and less likely to happen to us than to happen to someone else (Bauman & Siegel, 1987; DeJoy, 1989; Dunning, 1993; W. B. Hansen & Malotte, 1986; D. M. Harris & Guten, 1979; Linville, Fischer, & Fischhoff, 1993; Perloff & Fetzer, 1986; Robertson, 1977; Svenson, 1981; Zakay, 1983, 1984).

As already noted, automobile drivers tend, on average, to consider themselves to be safer than the average driver (Svenson, 1981; Svenson, Fischhoff,

& MacGregor, 1985); they also consider their chances of being involved in an accident to be lower as a driver than as a passenger (Greening & Chandler, 1997; McKenna, 1993; McKenna, Stanier, & Lewis, 1991). More generally, people tend to be more optimistic about events they perceive to be under their control than about those they consider not to be (Budescu & Bruderman, 1995; DeJoy, 1989; P. Harris, 1996; Hoorens & Buunk, 1993; Zakay, 1984). People launching new businesses typically estimate their chances of success to be high, despite evidence that the majority of small businesses fail within a few years of establishment (Cooper, Woo, & Dunkelberg, 1988). People sometimes discount the seriousness of a medical risk if they have reason to believe themselves to be especially susceptible to it (Ditto, Jemmott, & Darley, 1988; Ditto & Lopez, 1992; Jemmott, Ditto, & Croyle, 1986; Kunda, 1987). Such egocentric biases can also be affected by mood, so that people are likely to see a rosier future for themselves, and others, when in a happy frame of mind than when in a sad one (W. F. Wright & Bower, 1992).

Risk Perception Versus Risk Acceptability

Some people willingly, even enthusiastically, engage in risky behavior. According to Press (1975) about one third of the U.S. population lives in the two regions where the risk of major earthquakes is greatest, and people live there without taking any special precautions to prevent or minimize damage from earthquakes, should they occur. Of course, many of the people who live in areas where the risk of earthquakes is relatively high may be unaware of the nature of this risk or, if they are aware of it, may feel unable to do much about it. But some people also attempt to cross oceans in rowboats, go over Niagara Falls in barrels, drive cars in races at excessively high speeds, or voluntarily engage in other types of behavior for which the high risk can hardly be in doubt.

The distinction between engaging in risky behavior in the full realization of the risks involved and engaging in that behavior in ignorance of, or having grossly underestimated, the risks is one of some practical significance. Consider, for example, the case of drivers who engage in risky behavior—by driving too fast, driving while drinking, following leading vehicles too closely, passing with insufficient forward vision, failing to use seat belts, driving unsafe vehicles, ... How one would go about trying to modify the risky behavior in any particular case would depend on whether one assumes that the driver is unaware of the magnitude of the risk that is being taken or that the driver is fully aware of the risk that is being taken and is taking it willingly. The first possibility calls for finding a way to make the driver aware of the risk that is being taken; the second requires something more than provision of this knowledge, which the driver already has.

Factors Affecting the Acceptability of Risks

Hallman and Wandersman (1992; see also Wandersman and Hallman, 1993), who have taken the position that people's responses to environmental threats are predictable and not as irrational as they sometimes appear to be, note that those responses are determined not only by the perceived likelihood of a threat being realized, but also by a variety of other factors. People are likely, for example, to find a risk that is voluntarily taken to be more acceptable than a risk of equal magnitude that is involuntarily imposed (Fischhoff, Slovic, & Lichtenstein, 1978; Starr, 1972). Risks that are under individual control, fairly distributed among a population, natural, familiar, detectable, well understood by science, and ethical are more acceptable than risks that are under governmental control, unfairly distributed, artificial, exotic, undetectable, or unethical.

Responses to perceived risk appear to be affected by personal values (Brody, 1984; Office of Technology Assessment, 1987), which, in turn are influenced by the social and cultural contexts in which people live (Bradbury, 1989; Covello & B. B. Johnson, 1987; Vaughn & Nordenstam, 1991). Citing the distinction made by Hance et al. (1988) between *hazard* factors, which are those usually measured objectively if possible, and *outrage* factors, which tend to be more social, political, or ethical in nature, Wandersman and Hallman (1993) suggest that people are likely to be concerned (or outraged) by a risk surrounded by outrage factors even if its probability of realization is low. Other investigators who have emphasized the predictability of the perception of and response to risk include Vlek and Stallen (1980), Baum, Fleming, and Singer (1983), and Cvetkovich and Earle (1992).

When assessing the acceptability of future, especially far-future, risks, it seems appropriate to discount them to some degree just because they are in the future, and many things can happen to modify or nullify them—one has no guarantee, after all, of even being alive at any future date. The evidence suggests that people do discount the future in the sense of attaching greater importance to present costs and benefits than to costs and benefits that could be realized at a future time. Possible future calamities are perceived as less serious the further in the future their possible occurrence is considered to be. Although it is hard to quantify the problem, it appears that people find it easy to overdiscount the future. Thus they satisfy immediate wants by incurring future debt that they are unable to pay, they smoke and use other addictive drugs with apparent lack of concern for the long-term risks to their health, and they often fail to incur modest immediate costs (health check-ups, insurance) in the interest of protecting against serious future problems (Klatzky & Messick, 1995; Platt, 1973).

In the context of economic decision making, there appears to be a general preference for risk aversion, the prevalence of which Kahneman and Tversky

(1979b) have referred to as "perhaps the best known generalization regarding risky choices" (p. 263). Risk aversion appears to hold generally, especially when one's choice is between positive alternatives, but risk seeking is often observed when the alternatives between which one must choose are both negative, as, for example, when one's choice is between a sure loss and a gamble involving a possible loss of greater magnitude and a possible gain (Fishburn & Kochenberger, 1979; Markowitz, 1952; Shafir & Tversky, 1995). Kahneman and Tversky (1979b) attribute the preference for risk aversion in the domain of gains and for risk seeking in the domain of losses to the same psychological principle—the overweighting of certainty. However, it is not just that certainty is desirable, they suggest, but that certainty increases the aversiveness of losses and the desirability of gains.

Risk aversion for choices among positive alternatives and risk seeking for choices among negative alternatives are both incorporated in "prospect theory," developed by Kahneman and Tversky (1979b). According to this theory, the decision maker's focus is on the gains and losses that could be realized as a consequence of the decision and not on the decision maker's final state of wealth; and losses of specific amounts have a greater effect on subjective value than do gains of the same amount. This aspect of the theory might be accounted for by the fact that a gain of a specific amount is a smaller proportion of one's resulting total wealth than is a loss of the same amount: A gain of $25 on $100 is one fifth of one's resulting total of $125, whereas a loss of $25 from $100 is one third of one's remaining total of $75.

Buffon, in his "Essai d'Arithmétique Morale," published in 1777, also argues that a loss of a given amount is perceived as larger than a gain of the same amount: If one stakes half one's wealth on a gamble, one stands to increase one's wealth by one third and to decrease it by one half. Curiously, to justify the one third in this argument, he divides the gain by the resulting total wealth, but to justify the one half, he divides the loss by the original wealth. As Todhunter (1865/2001) points out, the argument would be even stronger if he divided by the resulting wealth in both cases (as do Kahneman and Tversky).

Prospect theory has proved to be predictive of behavior in many probabilistic choice situations (Kahneman & Tversky, 1979b; Shafir & Tversky, 1995). The asymmetric relationship between the subjective values of gains and losses is known as *loss aversion*. This relationship has important implications in a variety of contexts. Shafir and Tversky (1995) point out, for example, that a mediator of a dispute could increase the chances of obtaining an agreement by framing concessions as bargaining chips rather than as losses.

On the average, groups are inclined to take riskier decisions than individuals, and individuals are likely to be willing to accept greater risk when participating in a group decision-making process than when making decisions on their own (Bem, Wallach, & Kogan, 1965; R. D. Clark, 1971; Kogan & Wallach, 1967; Wallach & Kogan, 1965). This finding has been referred to as the "risky-shift" phenomenon, and may be a special case of participation in group discussion moving people to more extreme positions than they otherwise would express (Myers & Lamm, 1976).

On the basis of several studies of risk-taking behavior, Slovic (1972) has concluded that a propensity for risk taking is probably not a situation-independent character trait. One's tendency to take risks in one situation has proved not to be a good predictor of one's willingness to do so in qualitatively different situations. Whatever the explanation for different propensities for risk taking, it can be argued that from the point of view of rationality, one's attitude toward risk must be taken as a given. Allais (1979/1990) puts it this way: "It is accepted that a rational individual's scale of psychological values may differ from the monetary scale, and that he may have a greater or lesser propensity for safety or for risk. There seems to be agreement that this is an issue of psychology and not of 'rationality'" (p. 121).

Risk Communication

Inasmuch as both governmental policies and personal decisions often are made in response to perceived risks, it is of some importance that reliable data regarding risks be communicated in a way that will be correctly understood by the policy formulators and decision makers. In view of the considerable evidence that many people have difficulty in dealing with probabilistic concepts, finding effective ways to present risks to the general public is a challenge, and recognized as such (Allen, 1987; Hance et al., 1988; National Research Council, 1989). The point is made by reference to the Three Mile Island incident. The information immediately forthcoming from various sources was inconsistent and confusing (Goldsteen, Schorr, & Goldsteen, 1989). Lack of confidence in communication from government officials is indicated by the flight of about 200,000 people from the area following the issuing of a gubernatorial advisory that pregnant women and preschool children living within 5 miles of the plant (an estimated total of about 3,500 people) might want to evacuate and that everyone living within 10 miles should consider staying indoors (Erikson, 1990). On the other hand, overreaction is not the invariable response to risk. Many people refused to leave the vicinity of Mt. Saint Helens before its eruption in May 1980 (Saarinen, 1980), and people do elect to live in areas where risks of natural disasters are relatively high.

Several investigators have attempted to compare exposure to each of a variety of risks on a common metric, such as amount of decrease in one's life expectancy (B. Cohen & Lee, 1979; Sowby, 1965; R. Wilson, 1979). Slovic, Fischoff, and Lichtenstein (1981/1986) note that whereas such comparisons may serve a useful educational purpose, they fail to capture the complexity of people's attitudes toward risks that are affected by numerous variables, such as the degree to which a risk is controllable or avoidable, the potential it has to assume catastrophic proportions, and the threat it poses to future generations. The same investigators take the position that appropriate presentations of factual material within a comparative framework can do much to counter misperceptions and misestimates and to put risks in a more accurate perspective. People are more likely to see data regarding risks to be relevant to them personally if the data are framed in terms of personal risk than if presented as population statistics (L. Jeffrey, 1989; Sharlin, 1986).

Need for Better Understanding of Risk Assessment, Communication, and Perception

A better understanding of how risks are assessed and communicated is of some practical urgency, because policy decisions are often made for the explicit purpose of decreasing the risks of future catastrophes. To the extent that the actual risks differ from what they are perceived to be, actions that are taken in the interest of avoiding or decreasing them may be ineffective or dysfunctional (Gould et al., 1988). Our ability to anticipate risks associated with technological developments of various sorts in the past has not proved to be great; the need to do better in this regard becomes ever more urgent as the potential of technology for both good and ill continues to increase.

Of special concern is the question of how to assess risks of events for which frequency data do not exist. How, for example, does one assign a probability to the possibility of the accidental development and release of a lethal virus from genetic experimentation? Or to the intentional contamination of a city's water supply by a terrorist organization? Or to the finding of a cure for cancer? There seems to be little alternative to reliance on the opinions of experts, and this can be disconcerting when experts' opinions differ greatly on what the probabilities are. We can require that experts make the basis of their opinions explicit, and usually they are more than willing to do so in some detail. It is up to us then to judge which of the rationales that are put forth in defense of various positions that are taken that we find most persuasive. Making this judgment can be very difficult, of course, if we lack the background knowledge that is necessary to understand the rationales that are offered.

Probabilities can be represented graphically in a variety of ways. Unfortunately none of these representations appears to be able to compensate for a lack of familiarity with probability (Ibrekk & M. G. Morgan, 1987). People who have not had a nontrivial amount of exposure to probabilistic concepts seem to find it difficult if not impossible to think in these terms. Even introductory college courses in statistics and probability do not always suffice to ensure that people who have completed them will do appreciably better than people who have not on problems that require probabilistic reasoning many years after their training.

When we are dealing with the problem of understanding risks for the purpose of setting public policy, there is a responsibility, I want to argue, both on the part of experts to make the rationales for their opinions as accessible to a lay public as they can, and on the part of lay people to make an effort to obtain the knowledge that is necessary to understand the issues. When experts hide their rationales for their opinions behind technical jargon that is accessible only to their colleagues, if to them, they do not deserve to be taken seriously by the general public. On the other hand, it is not clear that one who is unwilling to make any effort to understand an issue can legitimately claim a right to have an opinion on it.

Because the same risk can appear to be different when reported in different ways (Slovic, 1987; Tversky & Kahneman, 1981), the question of how best to report risks is an important one. In the absence of a demonstration that a particular method of reporting invariably promotes better understanding than do alternative methods, an approach that has been recommended is that of expressing a given risk in more than one way (R. Wilson & Crouch, 1987).

Do people want to understand the risks they face as individuals? One suspects that this question has a complicated answer and that we do not know yet what it is. There is some evidence that, at least in the context of health and medical problems, many people desire reliable information about their risks (Alfidi, 1971; Weinstein, 1979). On the other hand, there is evidence too suggesting that people may sometimes avoid medical examinations when they have reason to expect them to produce bad news (Klatzky & Messick, 1995). The discovery of the gene for Huntington's chorea made it possible to determine whether individuals have the gene, but not all who have reason to suspect that they might have it elect to find out for sure.

SELF-EVALUATION

> Many things depend on people's subjective assessment of what they know and do not know: whether they volunteer for certain roles or tasks, whether they seek further practice or instruction, and whether they instill confidence in others, as well as the answers they give to questions from superiors and subordinates and the affect they induce on others by facial expressions and body language. (Jacoby, Bjork, & Kelley, 1994, p. 57)

The ability to reflect on one's own thought processes is a fascinating ability and one that we do not understand at all well. We know that we have this ability, however, and we use it in a variety of ways. We can make judgments and then make judgments of those judgments. We can estimate the likelihood of a specified event and then rate our confidence in the estimate. We can assess what we know on a particular subject relative to what there is to know.

This is not to claim that we can do these things well or that the self-evaluation judgments we make are necessarily accurate. Such judgments must have an accuracy greater than chance, or they would not be worth making, but it would be surprising if they were highly accurate, given the introspective basis on which they are made. Some of these judgments may even be a bit paradoxical. Consider, for example, the problem of assessing how much one knows relative to what there is to know about a given subject. In order to make such an assessment, one must not only know what one knows about the subject, but in some sense one must also know what there is to know—which is to say one must know what one does not know—in order to make the comparison.

Psychologists have done numerous experiments in which people have been asked to express their degree of confidence in various types of judgments that they themselves have made. Interest in the topic goes back at least to the early decades of the 20th century (Hollingsworth, 1913a, 1913b; Lund, 1925; Trow, 1923). One purpose of confidence-judgment studies has been to investigate how confidence, or degree of certitude, relates to a variety of more objective variables, such as the probability that a judgment made with a given level of confidence is correct.

When participants in experiments have expressed confidence as probability estimates or as ratings that can be transformed—with some plausible assumptions—into probability estimates, it has often been possible to compare these probability estimates with performance on the primary task, determining for each confidence estimate what percentage of the judgments on the primary task to which that estimate was assigned were correct. Plots of actual percentage correct against percentage correct "predicted" by the confidence estimates have been referred to as calibration curves; perfect calibration is represented by the unit line, which indicates that for any given confidence estimate, *X*, the proportion of all the judgments with that estimate that were correct was *X*. Calibration curves have been plotted by numerous investigators for judgments obtained under a wide variety of conditions (J. F. Yates, 1990, 1994).

Several different measures of performance that can be derived from confidence estimates of the type used in calibration studies have been defined. Three are *over/underconfidence, calibration,* and *resolution.* The following definitions are from Lichtenstein and Fischhoff (1977):

$$\text{Over/underconfidence} = \frac{1}{N}\sum_{t=1}^{T} n_t\left(r_t - c_t\right),$$

where N is the total number of responses, n_t is the number of times the response r_t was used, c_t is the proportion correct for all items assigned probability r_t, and T is the total number of different response categories (confidence levels) used.

$$\text{Calibration} = \frac{1}{N}\sum_{t=1}^{T} n_t (r_t - c_t)^2 ,$$

$$\text{Resolution} = \frac{1}{N}\sum_{t=1}^{T} n_t (r_t - c)^2 ,$$

where c is the overall proportion correct.

As Lichtenstein and Fischhoff (1977) point out, Equation 1 is the difference between the mean confidence (expressed as the probability of being correct) and overall proportion correct; obviously, the smaller the number, the better. A problem with Equation 1 as an indication of calibration is that, because it uses signed differences, overconfidence for some items or persons can counterbalance underconfidence for others, thus giving an indication of accurate confidence that could be misleading. A possible answer to this problem, which has been proposed by Adams and Adams (1960) and Oskamp (1962), is to use the absolute values of the differences; another is Equation 2, which was proposed by Murphy (1973). Either gives an indication of how close the calibration curve is to the unit line (which represents perfect correspondence between confidence and performance); the latter weights larger differences proportionately more than the former. Equation 3, also proposed by Murphy, provides an indication of the degree to which differences in confidence are predictive of commensurate differences in proportion correct; it increases with the slope of the confidence-correctness function (being close to 0 when this is relatively flat) and it can be thought of as a measure of sensitivity.

A fourth measure also proposed by Murphy (1973, 1974) and mentioned by Lichtenstein and Fischhoff (1977),

$$\text{Knowledge} = c(1 - c),$$

contains no representation of confidence, but provides a measure of correctness that is of interest primarily because, when combined with the measure of calibration defined by Equation 2 and the negation of the resolution score defined by Equation 3 it yields one of a class of what have been identified as proper, or admissible, scoring rules" (Shuford, Albert, & Massengill, 1966), discussed briefly in chapter 10. We should note that the preceding equation is 0 when either all the answers are correct or none of them is and increases as the proportion correct gets closer to 0.5.

Overconfidence

> We're so often cocksure of our decisions, actions, and beliefs because we fail to look for counterexamples, pay no attention to alternative views and their consequences, distort our memories and the evidence, and are seduced by our own explanatory schemes. (Paulos, 1998, p. 55)

Probably the broadest generalization that comes out of calibration studies and other investigations of judgments of judgments is that people tend to express a higher degree of confidence in their judgments than is justified by the accuracy of those judgments when they involve quantitative variables such as probabilities. Calibration studies have generally shown overconfidence to be a more common failing than underconfidence (Alpert & Raiffa, 1982; Arkes & Harkness, 1983; J. J. Christensen-Szalanski & Bushyhead, 1981; Einhorn, 1980; Einhorn & Hogarth, 1978; Fischhoff, 1982; Fischhoff & Slovic, 1980; Kelley & Lindsay, 1993; Lichtenstein & Fischhoff, 1977; M. G. Morgan & Henrion, 1990; Paese & Feuer, 1991; Pitz,1974; Ronis & F. Yates, 1987; Slovik, Fischhoff, & Lichtenstein, 1977; Vallone, Griffin, Lin, & L. Ross, 1990). Reviews include Lichtenstein, Fischhoff, and Phillips (1982), Wallsten and Budescu (1983), O'Connor (1989), and Keren (1991). The finding that overconfidence in one's judgments is the rule has serious practical implications because, as Baron (1998) has pointed out, people are more likely to take extreme actions on the basis of views in which they have great confidence than on the basis of those for which they have reservations.

Much of the evidence of overconfidence comes from experiments in which people have answered general-knowledge questions—questions pertaining to the type of information found in almanacs or books of facts—with a forced-choice format and have, for each answer, expressed their confidence that it was correct, but overconfidence has been observed in other contexts as well. Retrospective judgments of comprehension of expository text have been observed to be higher than justified (Glenberg & Epstein, 1985; Glenberg, Sanocki, Epstein, & Morris, 1987; Glenberg, Wilkinson, & Epstein, 1982), as have expressions of confidence in predictions of future events over which one is assumed to have some degree of control (S. J. Hoch, 1985). Fischhoff and Slovic (1980) asked people to perform a variety of difficult or nearly impossible tasks—identification of the nationalities of people on the basis of samples of their handwriting, projection of stock price activity, prediction of horse race outcomes—and found that people's performance typically was not as good as their confidence indicated they thought it would be. Griffin, Dunning, and L. Ross (1990) showed that people tend to be overconfident of their ability to predict either their own behavior or that of others in specified situations. People's assessments of how well they know some subject matter appear not to be very indicative of how well they will do on an exam on the subject (Cull & Zechmeister, 1994; Mazzoni & Cornoldi, 1993).

Experts have also often been found to be overconfident when assessing their own knowledge in their areas of expertise. When asked to predict the outcomes of their own cases, attorneys, for example, appear to be overconfident in the aggregate, in the sense of being apt to express confidence of obtaining outcomes better than those they actually obtain (E. F. Loftus & Wagenaar, 1988; Wagenaar & Keren, 1986). Other professionals who have been found to be overconfident when making judgments in their own areas of expertise include physicians (Faust, Hart, & Guilmette, 1988; Lusted, 1977), psychologists (Oskamp, 1965), and engineers (Kidd, 1970). Experts appear to do better when there is a reliable basis for statistical prediction, such as when predicting bridge hands (Keren, 1987) or betting odds for horse racing (Hausch, Ziemba, & Rubenstein, 1981).

People, including experts, also tend to be overconfident of their ability to estimate failure rates of system components. When asked to specify one failure rate that they would expect only 5% of components to exceed and another that they would expect only 5% of components to fall below, they typically pick values that are insufficiently extreme (Lichtenstein, Fischhoff, & Phillips, 1977). Kahneman and Tversky (1973) refer to the confidence that people feel for highly fallible judgments as the "illusion of validity." Apparently experts are not immune from this illusion (Nuclear Regulatory Commission, 1978; Slovic, Fischhoff, & Lichtenstein, 1981).

Kuhn, Weinstock, and Flaton (1994) have demonstrated a connection between confidence in a decision and the approach that one takes in reaching it. They found that some participants (satisficers) in a mock jury trial tended to use evidence selectively in order to build up a single view of what had happened, whereas others (theory-evidence coordinators) considered more than one possibility and attempted to evaluate all of them in light of the accumulating evidence. A particularly thought-provoking aspect of their results is the fact that those participants who took the former approach had greater confidence in their decisions than did those who took the latter.

Lichtenstein and Fischhoff (1977) found the tendency toward overconfidence to be especially pronounced when people made judgments regarding which there was little objective basis for accuracy (e.g., whether drawings were made by Asian or European children) and for which accuracy was close to chance. Calibration and resolution in this case were also poor. The same investigators found that providing people with some task relevant knowledge decreased overconfidence and improved calibration and resolution, but that when knowledge was very high (so that more than about 80% of the questions were answered correctly) underconfidence sometimes became the rule.

They concluded that it is not the case that people who know more know more about how much they know, in general, but rather that confidence appears to be most predictive of what one knows at intermediate levels of knowledge; when one knows very little, confidence is likely to be too high, whereas when one knows a lot (gives correct answers on a large majority of the questions) it is sometimes moderately too low. Their results over a series of experiments seem also to be consistent with the conclusion that resolution is somewhat less affected by changes in the difficulty of the task than are under or overconfidence or calibration. Making the question-answering task easier or more difficult, for example, appears to have the effect of moving the entire correctness-confidence curve up or down rather than changing its slope.

One account of why people tend generally to be overconfident in their answers is that once having identified a plausible answer to a question or scenario for the future, they fail to consider possible alternative answers or scenarios (Griffin et al., 1990; S. J. Hoch, 1985; Shaklee & Fischhoff, 1982). A closely related hypothesis is that when one has produced a tentative answer to a question, people find it easier to bring to mind evidence in support of that answer than evidence against it (Graesser & Hemphill, 1991; Koriat, Lichtenstein, & Fischhoff, 1980). This account links the tendency to be overconfident of one's answers with the idea of a pervasive confirmation bias (Nickerson, 1998). The bias is expressed in this case as a tendency to bring to mind information that will confirm the answer one has produced and to overlook information that would count against it. This is similar to the explanation suggested by Nisbett and L. Ross (1980) of why people may persevere with a belief even after learning that the information on which the belief was based was fictitious: after having formed the belief they sought and found independent data to corroborate it.

Another suggestion, not necessarily in conflict with this one, is that a major factor contributing to confidence in an answer is the speed (Nelson & Narens, 1990) and/or ease (Kelley & Lindsay, 1993) with which the answer comes to mind. Nelson and Narens found a correlation between confidence and speed of answering even when the answers were incorrect, although for a given speed, confidence was higher for correct than for incorrect answers. V. L. Smith and H. H. Clark (1993) also found that when people were able to answer a question, the higher the confidence in the answer, the more quickly it was produced.

Kelley and Lindsay (1993) manipulated ease of access by exposing potential answers, some correct and some incorrect (but plausible), to questions before asking the questions. Prior exposure increased the speed with which participants produced those answers and their confidence in them when they used them, in both cases. This was true even in a condition in which none of the items on a preexposed list was a correct answer to the subsequent question and participants were informed of this and told to ignore the list during the question-answering phase of the experiment. The fact that they appeared to be unable to avoid the influence of prior exposure on feelings of confidence was taken as strong evidence that confidence is based in part on the ease with which answers, right or wrong, come to mind.

Griffin and Tversky (1992) make a distinction between strength (extremeness) of evidence and weight (predictive validity) of evidence and hypothesize that people tend to focus primarily on the former and make some—but typically insufficient—adjustments in response to the latter. This hypothesis leads to the expectation that overconfidence will be the rule when strength is high and weight low, whereas underconfidence will prevail when the opposite is the case. Griffin and Tversky argue that this hypothesis can reconcile the apparent discrepancy between the finding of conservatism in the updating of posterior probabilities in Bayesian decision making (Edwards, 1968) and the finding that people often make radical inferences on the basis of small samples (Tversky & Kahneman, 1971). Conservatism or underconfidence has typically been found, they suggest, when people have been exposed to large samples of data of moderate strength, and radicalism or overconfidence has generally been observed in situations involving moderately strong effects based on small samples; so both phenomena follow, according to this view, from the dominance of evidentiary strength over weight.

Numerous studies have shown that medical diagnoses based on case statistics tend to be more accurate than those based on clinical judgments, and that despite this fact clinicians typically have greater confidence in their own judgments than in those derived statistically from incidence data (Dawes, 1976). Studies yielding predictions from case statistics that are at least as good as, and usually better than, those from judgments of experts include those of Dawes (1971), Wiggins (1981), Wedding (1983), Leli and Filskov (1984), Meehl (1986), among many others. A simple formula developed by L. R. Goldberg (1965, 1968) distinguishes neurotics from psychotics on the basis of Minnesota Multiphasic Personality Inventory scores with 70% accuracy. The fact that 70% is greater accuracy than any group of clinicians had attained on this task did not suffice to convince clinicians to use the formula instead of their own individualized approaches (Arkes et al., 1986). One study of the way clinicians score projective tests showed 18.5% of them using standard procedures and 81.5% using their own individualized procedures (Wade & T. B. Baker, 1977).

Although the quality of decision making has not always been found to increase with the quantity of relevant information available to the decision maker (L. R. Goldberg, 1968; Hayes, 1964), confidence may increase with the amount of information available even when accuracy does not. Clinicians' confidence in their judgments about cases, for example, has been observed to increase with the amount of information on which the judgments were made even when the accuracy of the judgments did not improve (Oskamp, 1965; Ryback, 1967).

Allwood and Granhag (1996) found that the accuracy or realism of confidence judgments improved when people were asked to evaluate the extent of their knowledge in the area of a question they were about to be asked. These investigators interpreted this result as at least suggestive evidence that people's evaluations made them more aware than they otherwise might have been of the limitations of their knowledge in the specified areas. An interesting aspect of Allwood and Granhag's results is that participants judged their knowledge to be more nearly comprehensive for broadly defined domains than for narrowly defined ones.

Confidence, or degree of certainty, has usually been expressed in calibration studies as a rating on a linear scale (e.g., a 7-point scale anchored at one end by "certainty" and at the other by "pure guess"). Sometimes, however, people have been asked to estimate odds, the ratio of the chance that a particular situation pertains to the chance that it does not. The most extreme examples of overconfidence have been found in these cases. Fischhoff, Slovic, and Lichtenstein (1977) had people judge which of each of several pairs of causes of death was the more prevalent, and express their confidence in each judgment in terms of the odds that it was correct. Subjects tended to give odds that were way out of line with their primary task performance especially with the easiest comparisons: Odds given for comparisons on which the participants were right 90% to 95% of the time ranged from 10,000:1 to 1,000,000:1. One possible explanation for such unrealistic numbers is that the participants lacked a clear understanding of what "odds" means. A lecture on probability and odds caused a reduction in the use of such extremes but did not eliminate it. Slovic et al. (1977) also found that people tended to be overconfident when asked to judge the odds that their answers to general-knowledge questions were correct. It appears that overconfidence is the rule in calibration studies independently of whether confidence is expressed with probabilities or with odds (Hazard & C. R. Peterson, 1973; Seaver, von Winterfeldt, & Edwards, 1978).

Reducing Overconfidence

Investigators have explored various ways of trying to reduce the overconfidence that people express in their judgments (Arkes, Christensen, Lai, & Blumer, 1987; Griffin et al., 1990; May, 1986; Sniezek, Paese, & Switzer, 1990). With a two-alternative choice task, Koriat et al., (1980) got people to make confidence judgments that were more appropriate for their performance by having them list reasons for and against each alternative before choosing their answer and expressing their confidence in it. The tendency toward overconfidence was also diminished by having people identify an argument against their answer, after choosing the answer but before making the confidence judgment. Other studies in which people have been asked to evaluate or justify their views, especially when the evaluation includes providing reasons against one's own position, have yielded a reduction in confidence (Fischhoff, 1977; S. J. Hoch, 1984, 1985; Tetlock & Kim, 1987). Being forced to consider multiple construals of situational details has also been shown to reduce overconfidence in one's predictions of one's own future behavior in specified situations (Griffin et al., 1990).

Although such results encourage the belief that the tendency to be overconfident can be countermanded, to some degree, numerous attempts to reduce overconfidence that have met with very limited, if any, success testify to the persistence of the tendency and the difficulty of reducing it (Ferrell & McGoey, 1980; Fischer, 1982; Fischhoff & MacGregor, 1982; Seaver et al., 1978). Simply informing people of the tendency and asking them to avoid it appears not to work (Lichtenstein & Fischhoff, 1980). Lichtenstein and Fischhoff demonstrated that training-induced improvements in calibration may fail to generalize strongly to probability assessment tasks different from those used in the training situation.

Some data suggest that professional weather forecasters tend to be well calibrated, at least with respect to their weather predictions (Murphy & Winkler, 1971, 1974, 1977; Winkler & Murphy, 1968) and this has been attributed to the fact that they receive relatively constant and immediate feedback regarding the accuracy of their predictions, which is the kind of information that makes learning feasible. It should be noted too, however, that weather forecasters typically do not express confidence in their judgments about the weather, they make probabilistic predictions about the weather per se. That is they do not typically say "It is going to rain tomorrow, and my confidence is .9 that that prediction is correct"; they are more likely to say "The probability that it will rain tomorrow is .9." One might argue that these amount to the same thing, but not everyone will necessarily be convinced that that is the case.

There is at least suggestive evidence that overconfidence in judgments tends to be less when the judgments are accompanied by choices or consequential decisions than when they are simply expressions of opinions on which no action is taken (Kahneman & Lovallo, 1993; Paese & Sniezek, 1991). There are indications also that the poor calibration that some studies have reported could have been due, at least in part, to the use of written materials that were appropriate for a reading comprehension level beyond that of the participants (C. A. Weaver & Bryant, 1995; Weaver, Bryant, & Burns, 1995).

Bases for Overconfidence

Studies of Bayesian decision making have shown that people often overestimate the completeness of the sets of hypotheses provided about the possible states of the world (Fischhoff et al., 1978; Gettys, S. D. Fisher, & Mehle, 1978; Mehle, Gettys, Manning, Baca, & Fisher, 1981). Given this tendency, the use of previously prepared check lists in decision-making situations would appear to have some merit, and there is evidence that decision making can be improved in this way (de Dombal, Leaper, Horrocks, Staniland, & McCann, 1974). Users of electronic document searches often overestimate, sometimes by large amounts, the percentage of the relevant documents that a search has returned (Blair & Maron, 1985; MacGregor, Fischhoff, & Blackshaw, 1987). These and similar results are suggestive of a general tendency to overestimate the completeness of lists that are presented to us for various purposes.

Such a tendency can help explain, at least in some cases, the common finding of overconfidence in judgments of various types. If I can think of only two or three possible hypotheses to account for some event and I believe my short list to be relatively complete, when in fact it is not, I might judge the probability that my favored hypothesis is correct to be quite high, on the assumption that it is one from a small number of possibilities; whereas if I realized that it was really one from a large number of possibilities, my confidence in its correctness might be considerably lower.

Gettys et al. (1978) have proposed essentially this account of the finding that people tend to be overconfident of the probable adequacy of stated hypotheses: People not only fail to generate some plausible hypotheses because of memory failure or lack of knowledge, but they judge the set of hypotheses they have before them (produced from memory or provided by someone else) to be more complete than it is because they overlook the limitations of their own knowledge and memory. In other words, they discount hypotheses they have been unable to generate by ignoring the possibility there may be some.

A tendency to overestimate the completeness of lists can be seen as itself a manifestation of overconfidence. In estimating any set to be more complete than it really is, one is, in effect, taking the fact that one cannot think of members of the set that are not included in the enumeration as evidence that they do not exist: If many do exist, one has overestimated either one's knowledge of the set or one's ability to retrieve from memory the information retained there. In either case, one has shown more confidence in one's cognitive resources than warranted.

Overconfidence in one's own judgments may be a consequence, in part, of the fact that judgments that are made outside the laboratory are typically made for the purpose of choosing among action alternatives, and such choices often have the effect of precluding the possibility of obtaining certain types of evidence that would show them to have been poor ones, if they were. Inasmuch as decisions often rule out the possibility of discovering what the consequences of different choices would have been, the decision maker may not get the kind of feedback that would be expected to shape a more accurate model of one's judgment ability over time.

Gigerenzer (1991b, 1994) argues that the "overconfidence bias," as described earlier, is not really a bias, at least as it relates to some interpretations of probability. Proponents of a frequentistic interpretation of probability, the dominant interpretation since the middle of the 19th century, would not, he suggests, recognize the meaningfulness of the application of probability to single events. Subjectivists, on the other hand, would accept the applicability of probability to individual events, but would not necessarily hold that the probability one assigns to an event should be determined by its relative frequency of occurrence. In short, according to Gigerenzer, "A discrepancy between confidence in single events and relative frequencies in the long run is not an error or a violation of probability theory from many experts' points of view. It only looks so from a narrow interpretation of probability that blurs the distinction between single events and frequencies fundamental to probability theory" (1991b, p. 89). From the point of view of a dyed-in-the-wool frequentist, inasmuch as probability theory is (only) about relative frequencies, and does not apply to single events, "no statement about confidences can violate the laws of probability" (1994, p. 153).

Gigerenzer (1991b; see also Allwood & Granhag, 1996; Keren, 1991) also reports results obtained by him and his colleagues that suggest that people do not equate level of confidence with expected relative frequency, despite the experimenter's instruction to do so. When they asked people to report their level of confidence in the correctness of each of many general-knowledge questions, they obtained the usual overconfidence result; when they asked the same people to estimate the percentage of several sets of such questions they had answered correctly, the overconfidence result was not obtained—in this case, the estimates were quite accurate. The way to make the overconfidence effect disappear, Gigerenzer argues, is to eliminate "the experimenter's normative confusion between single events and frequencies" (p. 90). Again, "the discrepancy between mean confidence and relative frequency of correct answers, known as 'overconfidence bias,' is not an error in probabilistic reasoning. It only looks that way from a narrow normative perspective, in which the distinction between single-event confidence and frequencies is blurred. If we ask our subjects about frequencies instead of single-event confidence we can make this stable phenomenon disappear" (Gigerenzer, 1993, p. 300).

This is a testable conjecture. Whether it holds for all observations of overconfidence is not clear. Even if it does, asking about frequencies is not the same as asking about confidence. And the question remains why people often express confidence that is higher than appears to be warranted by their performance. If degree of confidence does not map onto probability of being correct in a straightforward way, the question becomes: Exactly what does expressed confidence signify?

Confidence and Judgment Difficulty

One would like to believe that a specific level of expressed confidence in a judgment could be taken as a reliable indication of the probability that the judgment was correct, independently of the context in which the primary judgment and the confidence judgment were obtained. Confidence judgments would be more useful if their predictive significance was invariant across judgmental situations than if they meant one thing in one context and something else in another. Unfortunately, the evidence indicates that their significance does vary with situational factors; in particular, the degree of overconfidence that people express tends to increase with the difficulty of the questions asked; when the questions are very easy—as evidenced by a high percentage of correct answers on the primary task—overconfidence may cease to be the rule and underconfidence may be found (Lichtenstein & Fischhoff, 1977). This has been called the "hard–easy" effect.

In one study (Nickerson & McGoldrick, 1963), college students tried to identify which of four specified states of the United States is the largest in area. For each of a large number of such items, they also expressed their confidence on a 5-point scale, ranging from pure guess to certainty, that their answer was correct. Three sets of questions were used, representing three levels of difficulty. The easy set contained only items for which the largest of the four states was considerably larger than the next largest (at least 20 ordinal positions when states were ordered in terms of size); the difficult set contained only items for which the maximum difference between the largest and next largest state of any four was relatively small (not more than 5 ordinal positions); and the remaining set contained a mix of items from these sets plus many of intermediate difficulty.

The probability that an answer would be correct increased with the degree of confidence expressed in it; however the rate of increase was very slight for the difficult set of questions, greater for the easy set, and greatest for the mixed set. Of greatest relevance to this discussion is the fact that, as a predictor of performance on the size-judgment task, a given expression of confidence meant quite different things with the different sets. In particular the lowest confidence level ("pure guess") was associated with close-to-chance performance (about 29% correct) with the difficult set and with much better than chance performance (about 53% correct) with the easy set. Similarly, highest confidence ("certainty") was associated with about 34% correct with the difficult set and with about 76% correct with the easy set.

In a second study (Nickerson & McGoldrick, 1965) only the mixed set of items was used—to ensure a large range of difficulty—and the way in which confidence judgments were assigned by participants who performed well on the size-estimation task was compared with how they were assigned by those who performed poorly. For both groups the probability that an answer would be correct increased monotonically with the degree of confidence expressed in it; the rate of increase in performance with increased confidence was about the same for both groups, but for any given level of confidence the probability that the answer associated with it would be correct was considerably higher for the high performance group than for the low. If the confidence rating scale that was used is interpreted as an equal-interval scale anchored at 0 and 1 so the ratings can be interpreted as estimates of probability of being correct, both groups tended to be overconfident of their answers (answers were less likely to be correct on average than indicated by the probability estimates) and the low-performance group was considerably more overconfident than the high-performance group (members of this group tended to use lower confidence ratings, but not sufficiently much lower to reflect their much poorer primary-task performance).

One explanation that has been proposed of overconfidence in general and the hard–easy effect in particular starts with the assumption that people tend to be quite well calibrated in the sense of being good judges of their knowledge as it relates to situations they are likely to encounter in everyday life. Probabilistic judgments in everyday—"natural environment"—situations are based on cues that are generally effective indicants of the true state of affairs (the cue that one city is north of another is a generally effective indicant that it is probably colder than the other as well). Because of the nonrandom way in which general-knowledge questions are sampled for use in studies of calibration, the cues that work in real-world contexts may be misleading in the experimental situation and participants' reliance on them will result in an appearance of overconfidence in some cases (Juslin, 1993, 1994). Whether people are well calibrated with respect to real-life judgments is an empirical question that deserves further study.

Artifactual Contributions to Over- and Underconfidence

Juslin (1993) argues that if people are well calibrated to their natural environment, a minimum requirement for the observation of good calibration in the experimental situation is that the questions that people are to answer, and with respect to which they are to judge the probability of the correctness of their answers, are selected in such a way as to ensure that cues that are valid in the natural environment remain valid in the experimental situation. He contends that certain strategies that are commonly used to select items for use in experiments more or less ensure that the knowledge that is valid in participants' natural environment will be less valid in the experimental situation; more specifically, the argument is that such selection strategies typically result in sets of items for which cues leading to wrong answers are "overrepresented" relative to their commonness in the natural environment, and that this is especially so for relatively difficult items. The hard–easy effect can be an artifact, he suggests, of the partitioning of items into different difficulty categories after the fact on the basis of performance on the primary task. Supporting this explanation are the results of experiments showing good calibration for items selected at random from a set assumed to be representative of the natural environment (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1993, 1994). A counterexample is given, however, by Griffin and Tversky (1992). Also the Nickerson and McGoldrick (1963) study mentioned earlier, which showed the hard–easy effect, did not involve partitioning the items into different degree-of-difficulty categories on the basis of performance on the primary task.

A common finding in revision-of-opinion studies with a Bayesian paradigm was that people revise opinions on the basis of new data less than they should, according to the Bayesian prescription. They are said to be too conservative in their treatment of new data. As a consequence of this tendency, the (posterior) probabilities that people assign to events tend differ less from the prior probabilities than they should with application of Bayes's rule. One might say that the subjective posterior probabilities produced in these studies represent underconfidence. (It must be said that, from an alternative perspective, the finding might be interpreted as overconfidence in one's prior probability, or overconfidence in an existing opinion.)

Erev, Wallsten, and Budescu (1994) address this apparent conflict in the literature directly and demonstrate that it is possible to get evidence either of overconfidence or of underconfidence from the same data set, depending on how it is analyzed. They note that Bayesian experiments showing conservatism in the use of new data (underconfidence) have typically focused on subjective probabilities as a function of known objective probabilities, whereas calibration experiments have typically considered objective probabilities as a function of expressed subjective probabilities. They show that it is possible with a single data set to get the appearance of underconfidence with the first type of analysis and that of overconfidence with the second type (see also Erev, Bornstein, & Wallsten, 1993), and, with a log-odds model of how confidence in a judgment might be determined, they show that even one whose judgments are neither over- nor underconfident can appear to be either if the decision criterion includes a random error term.

Erev et al. (1994) do not conclude that all evidences of over- or underconfidence are artifactual. Such a conclusion is unwarranted, they argue, in view of the fact that the magnitudes of both over- and underconfidence can be varied systematically and that, for both types of analysis, the effects sometimes invert. Nevertheless, the demonstration that the same data set can be made to appear indicative of either overconfidence or underconfidence, depending only on how it is analyzed, is a powerful reminder of the need for care in the interpretation of experimental results. Erev, Wallsten, and Budescu argue that how one should look at judgment data depends on what is being asked: "Conditionalize on the event state (or on probability when it can be independently defined) when the research is aimed at understanding the underlying cognitive processes themselves. Conditionalize on response when the focus is on accuracy. In either case, the performance measure will not be fully understood without incorporating notions of error" (p. 526).

Contraction Bias

A common finding in calibration studies, when confidence is expressed or can be interpreted as estimates of probability of being correct, is that the lowest levels of confidence tend to be lower than the actual probability of being correct, whereas the highest levels tend to be higher than the actual probability of being correct (Adams & Adams, 1960; Fischoff et al., 1977; Lichtenstein et al., 1982; Nickerson & McGoldrick, 1963, 1965). This result may be seen as representing a type of contraction bias, suggesting, as it does, better performance than expected when confidence is low and poorer performance than expected when confidence is high.

People often display what has been called a contraction bias when judging physical magnitudes (Poulton, 1982), whereby they tend to overestimate relatively small magnitudes and underestimate relatively large ones. There is some evidence also of a sort of contraction bias involving judgments of the frequency of events, and especially of risky events, the tendency being to overestimate the frequency of the lower-frequency events and to underestimate that of the higher-frequency ones (Attneave, 1953; Braithwaite, 1975; Lichtenstein et al., 1978). Lichtenstein et al. found, for example, that when asked to judge the frequency of death from various causes, people tended to overestimate low-frequency causes and to underestimate high-frequency ones. Among the factors identified by the investigators as possible causes of consistent misestimates are differences in amount of exposure to specific causes, and differences in the memorability of particular events.

In all of these cases, there is the possibility of a statistical artifact. Consider confidence ratings. When people use extremely high confidence levels, say 1.0, any variability can only bring the actual probability of being correct down, because it is impossible to be more than 100% correct. In other words, it is difficult to do better than one expects when one expects one's performance to be nearly perfect. A slightly different, but similar, argument can be made with respect to the low end of the confidence scale. Consider all those judgments for which people have expressed minimal confidence, indicating that their performance on the primary task should be at a chance level. Suppose that people differ slightly in the strictness with which they apply a criterion for deciding that they have absolutely *no* knowledge that is relevant to the judgment and that their judgment is a *pure* guess. (This supposition could apply across people or to a given individual across judgments.) Any variability of this sort would be expected to result in judgments assigned minimal confidence to be correct with a probability somewhat greater than that expected by chance.

It may be easier to see this if one oversimplifies the situation and considers the judgments for which minimal confidence is expressed as falling into two categories: those that are truly pure guesses and those based on some amount of information, however small. The probability of correctness of the pure guesses should vary around chance, but that of the judgments in the other category should vary about something greater than chance; the two categories in combination should also be somewhat above chance.

Perhaps when the possibility of statistical artifacts is taken into account there remain nonartifactual results of the sorts described—low estimated probabilities underestimating performance, and high estimated probabilities overestimating performance and tendencies to overestimate the frequency of low-frequency events and to underestimate the frequency of high-frequency events, but before that conclusion can be drawn, the possible artifacts need to be given due consideration.

The Hyperprecision Effect

Not all studies of confidence have had people express confidence level by giving confidence or degree of certainty ratings or probability estimates. Pitz (1974) had college students indicate their degree of uncertainty about the population of each of 23 countries in the Americas in the following way. For each country a participant was to provide two numbers such that, in his or her opinion, the true population was equally likely to be (a) less than the smaller number, (b) greater than the larger, or (c) between the two. If participants had an accurate understanding of their own uncertainty about the populations involved, we would expect the true population to be in each of the "tertiles" about one third of the time. In fact on only about 16% of the trials did the true population fall between the two numbers given, which suggests that the participants typically believed they could bracket the true value more narrowly than they could in fact do so. Pitz referred to his result as the "hyperprecision effect."

In another study described by Pitz (1974) in which the tertile estimation procedure was used, the hyperprecision effect was also obtained when participants estimated populations or the heights of well-known buildings, but not when they estimated the ages of well-known people. In the last case, the center tertile included the true age 47% of the time, which is to say, people underestimated their ability to bracket these values, or overestimated their degree of uncertainty. Pitz attributed this result to the greater store of knowledge that people possess about people's ages. In fact the result does not really tell us that the participants knew more about ages than about populations and building heights; it tells us only that they knew more about ages, and less about populations and building heights, than they thought they did.

Pitz (1974) mentions also unpublished data obtained by Alpert and Raiffa showing a hyperprecision effect. In this case participants were asked to estimate the 1st and 99th percentile values for various quantities and gave numbers that typically bracketed the correct value about 50% of the time rather than 98%. In the aggregate, the results of these studies suggest that one is more likely to overestimate than to underestimate how precisely one can specify distribution percentiles, and this is consistent with the common finding of overconfidence in other types of judgment tasks.

SUMMARY

The abilities to estimate and to predict are valuable skills. The focus of this chapter has been on statistical estimation and prediction. This is a small piece of the broader topic that would include estimation, or prediction, of a variety of physical or temporal variables, but it is an important part inasmuch as many of the practical contexts in which the skill is needed are probabilistic in nature.

Research results suggest that people are reasonably good at estimating measures of central tendency of distributions and relative frequencies of events, and less good at predicting the time and resources that will be required to perform specified tasks. Estimates of time and costs typically err on the low side. One factor that appears to contribute to this problem is the ease with which people assume, perhaps unwittingly, that everything will go according to plan (overestimation of the probability of the conjunction of successes) and that nothing will go wrong (underestimation of the probability of the disjunction of possible problems).

Predictions of the outcomes of probabilistic events are subject to a number of misconceptions and biases. Perhaps the best known misconception is the gambler's fallacy. Predictions often are influenced by preferences—people tend to overestimate the likelihood of occurrence of what they wish to occur. Arguably this tendency makes lotteries appear more attractive than they might appear if people clearly understood the odds against winning. Predictions can also be influenced by intuitions about sample size and population size and how the two relate.

Opinions about the relative seriousness of various risks are shaped not only by the objective probabilities with which feared events occur, but also by irrelevant factors such as the publicity that events of different sorts receive. Also the acceptability of specific risks is determined not solely by the seriousness of the consequences of the occurrence of a feared event, but by factors such as one's sense of control of the risky situation (driver vs. passenger in a car), and whether exposure to the risk is voluntary or imposed. Determining how to communicate information about specific risks in such a way as to evoke appropriate action is an important research objective.

Evaluations of one's own estimation or prediction capability or performance frequently prove to be too generous, which is to say, people's expression of confidence in the accuracy of their own judgments often tend to be higher than their performance warrants. There is some evidence that people evaluate their own performance more accurately when they judge it in terms of percentage of answers correct than in terms of confidence on an answer-by-answer basis. Finding ways to improve people's ability to assess their own capabilities and performance more accurately is another continuing challenge for research.

CHAPTER

9

Perception of Covariation and Contingency

Estimates of the degree to which variations in two variables are coupled have been investigated under several topics, including covariation, co-occurrence, contingency, correlation and joint probabilities. Here I will use these terms as they were used in the studies cited, without making distinctions among them, but will tend to use covariation as the default generic term.

The ability to detect covariation is widely recognized to be an important one for any creature. As Alloy and Tabachnik (1984) put it, "Information about the relationships or covariations between events in the world provides people and animals with a means of explaining the past, controlling the present, and predicting the future, thereby maximizing the likelihood that they can obtain desired outcomes and avoid aversive ones" (p. 112). Several experimental tasks have been used to study the estimation of covariation. These include inspecting graphical representations of the relationship between two variables (correlation scatterplots), inspecting series of number pairs, inspecting full or partial two-by-two contingency tables, and looking for evidence of covariation in case records (e.g., looking for the coupling of smoking and lung cancer in medical records).

Some investigators have questioned whether people untutored in statistics have an abstract concept of contingency or correlation (L. J. Chapman & J. P.

Chapman, 1967; Crocker, 1981; H. M. Jenkins & W. C. Ward, 1965; Shaklee & Tucker, 1980; Shweder, 1977; Smedslund, 1963). Shweder (1977) suggests that people's estimates of co-occurrence in their own experience are more strongly dependent on resemblance or semantic similarity than on actual frequency of co-occurrence. He suggests that people not only base judgments of what goes with what on resemblance but that they seem unable to make use of frequency information or to reason correlationally. On the other hand, Nisbett and L. Ross (1980) note that, although laboratory studies indicate that people are not very good at detecting covariation, they nevertheless adapt well in social contexts that seem to require the ability to detect covariation with considerable accuracy. This, they suggest, requires an explanation. This does not invalidate the results of laboratory studies, but it does suggestion caution in extrapolating those results to real-world contexts.

Several investigators have studied the ability of people to estimate the correlation between series of number pairs and found it to be quite good (Erlick & Mills, 1967; Jennings, Amabile, & L. Ross, 1982). Such estimates tend to be more accurate when the correlations are positive than when they are negative (Erlick & Mills, 1967) and the errors that are made are typically in the direction of conservatism, which is to say in that of underestimating the strength of the relationship (Beach & Scopp, 1966; Jennings et al., 1982).

BIASED FOCUS

The relationship between two variables, A and B, is often represented by showing, in a table, the frequencies of occurrence of all four possible combinations of the presence or absence of each variable. An example of such a table, sometimes referred as a contingency table, is shown as Table 9.1.

Covariation in Contingency Tables

When people are asked to estimate the strength of the relationship between two variables based on information in two-by-two contingency tables, they often

TABLE 9.1

Illustrating the Tabular Representation of the Contingency Between *A* and *B*

	A Present	*A Absent*
B present	A and B	Not-A and B
B absent	A and not-B	Not-A and not-B

give too much weight to the cell representing the positive state of both variables (*A* and *B*) and pay too little attention to the other cells (Arkes & Harkness, 1983; Crocker, 1981; Doherty & Falgout, 1986; Einhorn & Hogarth, 1978; H. M. Jenkins & W. C. Ward, 1965; Kuhn, Phelps, & Walters, 1985; C. R. Peterson & Beach, 1967; Schustack & Sternberg, 1981; Shaklee & Mims, 1982; Shaklee & Tucker, 1980; Smedslund, 1963; Wasserman, Dorner, & Kao, 1990). The general finding is illustrated by a study by Smedslund in which a group of nurses reviewed a set of clinical cases in which a specific symptom sometimes co-occurred with a particular diagnosis. The set of cases selected for review included all four possible combinations (symptom-disease, symptom-no disease, no symptom-disease, no symptom-no disease) but the actual correlation between symptom and disease was zero. The nurses reported the correlation to be positive, however, on the strength of their observation that the symptom often co-occurred with the disease.

The cell of a two-by-two contingency table that tends to get the most attention—the one that represents cases in which both of the variables of interest occurred—is typically referred to as the "plus-plus" cell; the other three cells represent cases in which one or both of the variables failed to occur. With respect to this terminology, it is important to note that "plus-plus" need not refer to a specific cell (e.g., upper left) of a contingency matrix, but rather to the combination of variables on which the individual focuses. This is illustrated by an experiment by Crocker (1982) in which some participants were asked to judge whether practicing the day before a tennis match related to winning and others were asked to judge whether such practice related to losing. Both groups were asked which cells of the two-by-two contingency table that showed the frequencies of all combinations of practicing-not practicing and winning-losing would provide information useful for their task. The group focused on winning was most interested in the cell representing the combination of practicing and winning; the one focused on losing wanted to know about the cell representing the combination of practicing and losing. Both showed the plus-plus bias, but what constituted the plus-plus cell differed for the two groups.

Focusing on the plus-plus cell of a contingency table, while completely ignoring the other three cells, is an extreme form of biased focus and not very representative of behavior. What is more typical is an undue emphasis on the plus-plus cell and insufficient attention to the other cells. Attention typically has been unequally distributed among the remaining cells, with the two cells representing the presence of one or the other of the variables receiving more than the one representing the absence of both (Arkes & Harkness, 1983; Schustack & Sternberg, 1981; Shaklee & Mims, 1982). The strategy of focusing primarily on the plus-plus cell and paying insufficient attention to the others was noted by Inhelder and Piaget (1958) and interpreted by them as an indication of immature judgment.

None of this is to suggest that all the cells of a two-by-two contingency table are equally important and always deserving of equal weight in any attempt to judge the strength of relationship between two variables. How much weight one should give to the various cells of such a table depends, to some degree, on the assumptions on which one is working. J. R. Anderson (1990) shows by a mathematical analysis, for example, that the minus-minus cell should carry as much weight as the plus-plus cell only if it can be assumed that the prior probability of the effect in question occurring in the presence of the cause in question is equal to the prior probability of that effect not occurring in the absence of that cause. Thus, according to Anderson, the fact that people tend to focus on the plus-plus cell when attempting to determine the strength of relationship between two variables represented in a two-by-two contingency table is not compelling evidence that they are behaving irrationally; "the critical issue is what prior model is adopted in a rational analysis" (p. 160).

That being said, it must be noted that the consensus among investigators who have studied this situation appears to be that the degree to which people focus on the plus-plus cell and ignore or discount the others is not usually justified and typically leads to estimating the strength of relationship between two variables to be stronger than the data indicate it to be. Focusing on the plus-plus cell in two-by-two tables is sometimes seen as one of several examples of the difficulty people have in using disconfirming information.

How misleading a focus on the plus-plus cell can be depends on the nature of the relationship between the variables involved. If the variables are highly correlated, such a focus is not likely to lead one to conclude that they are not. It is easy to imagine circumstances, however, in which this focus could result in the conclusion of a strong relationship when one does not exist. This asymmetry helps account for the fact that this type of focus typically leads to an overestimation of the strength of the relationship between the variables involved.

It should be clear that by ignoring, or paying too little attention, to cells other than the plus-plus cell, one is extracting much less information about the possible relationship between *A* and *B* than a contingency table provides. Suppose that *A* is a necessary cause of *B* (or that *B* is a sufficient cause of *A*). In that case we would expect to find no occurrences of not-*A* and *B*. Or if *A* were a sufficient cause of *B* (or *B* a necessary cause of *A*), we would expect no occurrences of *A* and not-*B*. If *A* were a necessary *and* sufficient cause of *B* (or *B* a necessary and sufficient cause of *A*), we would expect to see entries only in the plus-plus and minus-minus cells. Conversely, if we observed one of these patterns, we would have grounds for suspecting a causal relationship of the associated type: if, for example, we observed a table with sizeable numbers in all cells except not-*A* and *B* and a zero in that one, we might begin to suspect that *A* was a necessary cause of *B*. None of these possibilities would be noted, however, if we focused only on the plus-plus cell.

Leaving aside the question of a possible causal relationship between *A* and *B*, the plus-plus cell by itself cannot tell us as much as there is to learn from the table about the degree to which the two variables covary. The higher the positive correlation between them, the more we expect to see the observations concentrated in the plus-plus and minus-minus cells; the higher the negative correlation, the more they would be concentrated in the two remaining cells. Inhelder and Piaget (1958) suggested that in judging the strength of contingency between two variables people may focus on the difference between (a) the sum of the plus-plus and minus-minus cells and (b) the sum of the remaining two cells: the larger the difference, the stronger the contingency. We should note that the difference can be either positive (indicating a positive correlation) or negative (indicating a negative one).

Covariation Among Events

However effective or ineffective people are at estimating correlations from two-by-two contingency tables, performance on this task tells us little about their ability to detect covariation among the events that may be represented in such tables. The little evidence there is on the ability to detect covariation among events suggests that at least when those events are arbitrarily paired stimuli such as number pairs, the ability to detect small to moderate correlations (less than about .6) is not very good (Jennings et al., 1982). It also appears that the tendency to ignore information represented by some of the cells of a contingency table is even greater when the information is presented sequentially and one must rely on memory to integrate it than when it is provided all at once in tabular form. More generally, people seem to be better at estimating contingency relationships when they receive the information in summary form than when they receive it serially on an instance-by-instance basis (W. D. Ward & H. M. Jenkins, 1965). If, when estimating the degree to which two events are correlated, people typically focus on instances in which the events co-occurred and overlook or disregard cases in which one occurred but not the other, they will overestimate the strength of the correlation. C. R. Peterson and Beach (1967) have suggested that this is in fact what people do.

The foregoing discussion has noted the tendency of people to do the equivalent of focusing on only one cell of a two-by-two contingency matrix when judging, either explicitly or implicitly, the strength of the relationship between two variables. Gilovich (1991) argues that this represents more than a bias that characterizes the behavior of individuals and that it reflects a tendency that can be observed in the media's reporting of news as well. More generally the claim is that various factors conspire to ensure that not all of the types of information represented by the cells of the contingency matrix are equally likely to come to our attention in the normal course of events, so it is incumbent on us, as individuals, to do some digging if we wish to get the full picture in specific instances.

EXPECTATIONS AND CAUSE–EFFECT ASSUMPTIONS

Expectation plays an important role in the perception of covariation, and has been identified by several investigators as a source of bias in covariation or contingency judgments (Crocker, 1981; Nisbett & L. Ross, 1980; C. R. Peterson, 1980). Covariation is especially difficult to detect when it occurs between variables that are not expected to be related; in contrast, when a relationship is expected, confirming instances may be given undue weight whereas disconfirming instances are overlooked or discounted (L. J. Chapman, 1967; L. J. Chapman & J. P. Chapman, 1967, 1969). Belief that a contingency exists can have the effect of increasing the chances that one will find evidence that tends to confirm the relationship and decrease the chances of obtaining evidence against it. This is one of many manifestations of a bias toward seeking confirming rather than disconfirming evidence of an existing belief (Nickerson, 1998).

That estimates of the degree of contingency between variables represented in a two-by-two table can be affected by the perceived causal connection between the variables was shown by Ajzen (1977), who varied the labels on tables and found that the more causally connected the two variables appeared to be, the higher the degree of estimated contingency for a given set of numbers. Other studies showing that the perceived degree of contingency can be changed by simply changing the labeling of the rows or columns of a table include those of Allan and H. M. Jenkins (1980) and Beyth-Marom (1982).

In a review of work on covariation detection, Alloy and Tabachnik (1984) reject the idea that people do not have an abstract concept of contingency, but concede that in the absence of certain mitigating factors they do frequently misjudge event relationships in systematic ways. They cite numerous experimental studies showing how biases might occur in making covariation judgments. They argue that people's assessment of covariation is influenced by both expectations and data or situational information. When expectations are strong and situational information weak, expectations exert the greater influence; when expectations are weak and situational information strong, the reverse is true; when both expectations and situational information are strong, either of two possibilities pertain. If expectations and situational information are consistent, one sees the covariation they both indicate; when they are inconsistent, expectations often prove to exert the greater influence. In the latter case, one is faced with what has been called a cognitive dilemma (Metalsky & Abramson, 1981) and evidence suggests that people tend to be biased in the direction of their initial expectations, but this tendency can be overridden if the situational information is sufficiently salient or compelling.

Alloy and Tabachnik (1984) summarize the conclusions that they believe can be drawn from empirical work on covariation perception this way:

> Perhaps, then, the most concise summary of the empirical work on covariation detection is that judgments of covariation are relatively accurate when people lack strong beliefs about the event relationship in question or when the situational information concerning the objective correlation between the events is congruent with people's preconceptions about the event relationship. When objective data and preconceptions are incongruent, judgments of covariation are frequently erroneous and biased in the direction of initial expectations. (p. 123)

This view has been challenged (Goddard & Allan, 1988) and defended (Alloy, 1988); it seems safe to say that precisely how good people are at detecting covariation and the conditions that determine the accuracy with which they do so remain subjects of debate.

ILLUSORY CORRELATION

> You can find perfect correlations that mean nothing for any three people and three characteristics, and in general for any *N* people and *N* characteristics.... Thus we tend to overestimate our general knowledge of others and are convinced of all sorts of associations (more complicated variants of "more shy, less intelligent") that are simply bogus. By failing to adjust downward our multiple correlation coefficients, so to speak, we convince ourselves that we know all manner of stuff that just isn't so. (Paulos, 1998, p. 27)

An illusory correlation is a perceived correlation between variables that are not correlated in fact; the term is also applied sometimes when the variables in question are correlated, but the perceived correlation is materially higher than the actual one (Allan & H. M. Jenkins, 1980; Crocker, 1981; D. L. Hamilton & Gifford, 1976; D. L. Hamilton & Sherman, 1989; H. M. Jenkins & W. C. Ward, 1965). Evidence suggests that people are likely to perceive correlations that do not exist when their prior expectations of such correlations are high (Camerer, 1988; L. J. Chapman & J. P. Chapman, 1967, 1969; Golding & Rorer, 1972; D. L. Hamilton, 1979). L. J. Chapman (1967) originally showed that the frequency with which words have been paired in experimental situations tends to be overestimated for pairs of words that have a strong associative relationship or are distinctive (e.g., unusually long). In the original study, participants also reported spurious correlations between features in clinical diagnoses of "patients" (e.g., suspiciousness) and features in person-drawings (e.g., peculiar eyes) the patients had made.

Nisbett and L. Ross (1980) summarize the Chapmans' work on illusory correlation by saying that "reported covariation was shown to reflect true covariation far *less* than it reflected theories or preconceptions of the nature of the associations that 'ought' to exist" (p. 97). The Chapmans stress the importance of the role of semantic associations as a source of beliefs about covariation and covariates.

Behavioral Patterns and Personality Traits

A form of stereotyping involves believing that certain unusual behaviors are more common among people who are members of a specific group than among those who are not. There is a perceived correlation between group membership and behavior. Such perceived correlations can be real, but they also can be illusory. One possible explanation of their occurrence is that unusual behavior by people in distinctive groups is more salient and easily remembered than similar behavior by people who are not members of those groups (Feldman, Camburn, & Gatti, 1986; D. L. Hamilton, Dugan, & Trolier, 1985). Illusory correlations can be the basis of stereotyping (D. L. Hamilton, 1981),

Estimates of the degree of correlation among personality traits appear sometimes to be based on the perceived similarity among the traits (Shweder & D'Andrade, 1980). Some assumed correlations may have little better basis than the belief that the variables involved should be correlated. One might believe on this basis, for example, that job satisfaction is highly correlated with job performance, but after reviewing more than 70 studies, Iaffaldano and Muchinsky (1985) concluded that the correlation between job satisfaction and job performance is only about .15.

Several investigators have been interested in the question of how accurately people estimate the consistency of human behavior, which is a form of covariation or correlation. Studies have produced mixed results, showing sometimes highly accurate estimates and at other times highly inaccurate ones. Expectations of high correlations among the behaviors of different individuals in similar situations or among behaviors of the same individual at different times in similar situations may be based on causal theories of behavior that assume that similar situations evoke similar response patterns, at least for the same person at different times and possibly for different people as well. The assumption regarding consistent behavior across people sometimes is limited to people with common personality traits or dispositions. Some investigators have argued that people typically overestimate the degree to which behavior in different situations can be predicted from trait variables. They would claim, for example, that the extent to which a friendly individual's behavior is consistently friendly or a hostile individual's behavior is consistently hostile is less than is generally believed. This misperception is known as the "illusion of consistency" (Jennings et al., 1982; Mischel, 1968; Mischel & Peake, 1982; D. R. Peterson, 1968).

Kunda and Nisbett (1986b) did a series of experiments on estimating behavioral consistency and got a wide range of outcomes. Their interpretation of their results emphasizes the role of three factors as determinants of accuracy of correlation estimates in the social world: (a) familiarity with the data, (b) codability of the data, and (c) whether the data to be correlated were drawn from distributions of the same kinds of events. The first two of these factors had been identified by other investigators; the third was noted for the first time as a result of Kunda and Nisbett's experiments.

As to why the third factor is important, Kunda and Nisbett (1986b) speculate that "correlations among variables coming from distributions of the same type are much easier to assess because in this case each pair of observations in and of itself contains information, namely, the distance between the two observations, that can be used to assess the correlation" (p. 218). Thus, for example, when two people are asked to evaluate the same academic course, using the same evaluation metric, the distance between their evaluations can be taken as some indication of the degree of correlation between the judgments. In contrast, if one is trying to estimate the degree of correlation between students' evaluation of a course and their performance in the course, the two evaluation metrics are not directly comparable, so an inference from the comparison to the correlation estimate is less straightforward.

Kunda and Nisbett (1986b) argue that this hypothesis helps account for the fact that people sometimes produce fairly accurate estimates of correlations among sets of numbers or among sets of readings of pointers on identical dials (Beach & Scopp, 1966; Erlick & Mills, 1967; J. C. Wright, 1962). And they suggest that when the three factors mentioned do not hold, people, including those trained in statistics, often greatly overestimate the degree to which an individual's typical social behavior can be predicted from a knowledge of that person's behavior on a given occasion.

Kareev (1995a, 1995b) notes that the sampling distribution of the most common measure of correlation, Pearson's *r*, is skewed when a correlation exists, and the more so the smaller the sample. As a consequence, the correlation one sees in a small sample is likely to be larger than that of the population from which the sample was drawn. Kareev argues that the limited capacity of working memory ensures that the samples people pay attention to will necessarily be small and therefore increase the chance that people will detect correlations that exist, at the expense of an increase also in false alarms. This amplifying effect of focusing on small samples could contribute to illusory correlation by making correlations appear to be larger than they actually are. Kareev notes the possibility that the working memory limitations of children may actually facilitate their detection of covariations and regularities that are the basis for the recognition of categories and causal relationships.

Cause–Effect Relationships

The phenomenon of illusory correlation is related to that of illusory cause–effect relationships. How one can come to believe in a cause–effect relationship that does not exist was nicely illustrated in an experiment by Schaffner (1985). People were asked to try to encourage a hypothetical student to arrive at school on time (8:30) by praising or criticizing the student's day-to-day behavior as appropriate. Participants typically praised the student when he arrived early and reprimanded him when he arrived late. As a result of the experience, most participants believed that reprimands had been more effective than praise in effecting the desired behavior.

In fact, the hypothetical student's behavior was independent of the positive and negative reinforcement provided by the participants, having been determined by computer before the experiment began. Because the predetermined arrival times varied symmetrically around 8:30, a late arrival was more likely to be followed by an earlier one than by a later one, and an early arrival was more likely to be followed by a later than an earlier one, simply as a matter of regression to the mean. Consequently if a participant invariably rewarded early arrival with praise and punished late arrival with a reprimand, the reprimands would more often be followed by improvement than would praise, thus leading subjects to conclude that reprimands were more effective than praise in bringing about the desired behavioral change.

PROBABILITIES OF CONJUNCTIVE OR DISJUNCTIVE EVENTS

By definition the probability of the joint occurrence of two independent events is the product of the probabilities of the individual events. If the probability of event A is $p(A)$ and that of event B is $p(B)$, the probability of the joint occurrence of A and B, usually written $p(AB)$, is the product of the individual probabilities, $p(A)p(B)$. The probability of the joint occurrence of two nonindependent events, A and B, is defined as

$$p(AB) = p(A)\,p(B \mid A)\,,$$

or as

$$p(AB) = p(B)\, p(A \mid B)\,,$$

where $p(B \mid A)$ is the probability of the occurrence of B, given the occurrence of A. (This formula is applicable to either independent or nonindependent events, because when A and B are independent $p(B \mid A) = p(B)$ and $p(A \mid B) = p(A)$.)

Overestimates of Conjunctions

People typically overestimate the probability of the joint occurrence of independent events relative to the given or estimated probabilities of the individual events (L. J. Cohen et al., 1972; Fleming, 1970; Wyer, 1974). This bias may be related to the tendency to base estimates of correlation between two events on the cases in which those events do co-occur and to pay too little attention to those in which they do not (C. R. Peterson & Beach, 1967).

There is some evidence that people also tend to overestimate the probability of the joint occurrence of nonindependent events. When, for example, law students were asked to estimate the probabilities of compound, not necessarily independent, events associated with fictitious criminal-case material, their estimates of $p(AB)$ were higher than they should have been relative to their estimates of $p(A)$ and $p(B \mid A)$ (Goldsmith, 1978).

Underestimates of Disjunctions

Overestimation of the probability of conjunctions of events (relative to the probabilities of the conjuncts) is logically similar to underestimation of the probability of disjunctions. In the first case, one's judgment of the probability that both of two events will occur is too high relative to the probabilities of each of the events separately; this is tantamount to underestimating the probability that at least one of the events will not occur. In the second case, one's judgment of the probability that at least one of two events will occur is too low relative to the probabilities of the events separately, and this is the same as overestimating the probability that both will not occur. Given several independent events of known probability, people typically underestimate the probability that at least one of them will occur (L. J. Cohen, Chesnick, & Haran, 1971, 1972; Tversky & Kahneman, 1974).

As noted in the preceding chapter, a tendency either to overestimate the probability of the joint occurrence of events or to underestimate the probability of the disjunction of events can have significant practical implications. Consider for example the problem of estimating the likelihood of all the steps in some sequential process being executed with no serious difficulties. The probability of the entire process being run off smoothly is the probability of the conjunction of each of the individual processes being executed without difficulty; the probability that the process will experience *some* difficulty is the probability of the disjunction of problematic individual steps.

In this example, both the tendency to overestimate the probability of a conjunction of events and the tendency to underestimate a disjunction lead to an overestimation of the probability that the entire process will be problem free, or conversely underestimation of the probability that it will experience some difficulty. These types of biases will cause trouble whenever the objective is to estimate the probability of success when success depends on the conjunction of two or more events. As noted in chapter 8, this may help account for why it is so easy to underestimate the time required to complete complex tasks (or the tendency to overestimate the probability of being able to finish such tasks in a specified time). We may simply assume that everything will go right and neglect to consider what may be a fairly high probability that at least one thing that will cause a delay will go wrong.

THE CONJUNCTION FALLACY

A special case of overestimation of the probability of the joint occurrence of events involves what has become known, variously, as the conjunction fallacy, the extension fallacy, or the compound-probability fallacy. It is axiomatic that the probability of a conjunction of two events can be no greater than the probability of the less probable of the individual events. Thus, if the probability of A is .5 and the probability of B is .4, the probability of the conjunction of A and B cannot exceed .4. This is analogous to the rule that the intersection of two or more sets can be no larger than the smallest of the sets involved. Although most people would presumably agree with the conjunction rule, as it is called, there are circumstances under which they tend to make judgments that violate it. When asked to judge the probabilities of conjunctions and the probabilities of their constituents, people sometimes judge the joint probability to be higher than the probability of the less probable constituent event (L. J. Cohen, Cheswick, & Haran, 1972; Kahneman & Tversky, 1982a; Slovic et al., 1976; J. F. Yates & Carlson, 1986).

Commonness of the Conjunction Fallacy

The conjunction fallacy is especially common, even among people trained in statistics and probability, when the conjunction is more "representative" of a class than is one of the constituents. Here is an example given by Kahneman and Tversky (1984) of how the fallacy works. The following personality sketch is read to a listener: "Bill is thirty-four years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities." Listeners are asked to rank order a set of statements about Bill from most to least probable. Among the set of statements used are the following three: (a) "Bill is an accountant"; (b) "Bill plays jazz for a hobby"; and (c) "Bill is an accountant who plays jazz for a hobby" (p.297). The description of Bill had been constructed so as to be representative of an accountant and unrepresentative of a person who plays jazz for a hobby. Most participants considered it more likely that Bill was an accountant than that he played jazz for a hobby and, in accordance with the conjunction fallacy, they considered it more likely that he was an accountant who played jazz for a hobby than that he played jazz for a hobby. That is to say, Sentence (c) was considered less likely to be true than (a), but more likely than (b).

This, of course, violates the rules of probability and is easily seen to be fallacious by viewing the problem as one of class membership. The class of all people who play jazz as a hobby includes, as a subclass, the class of all accountants who play jazz as a hobby. One is a member of the former (more inclusive) class by virtue of being a member of the latter, so the probability of being in the former class cannot be smaller than the probability of being in the latter. Osherson (1995) calls judging the latter probability to be smaller than the former an example of probabilistically incoherent reasoning and argues that the representativeness of category instances plays a key role in determining when it will occur.

The conjunction fallacy has been described as an instance of "scenario thinking" (Dawes, 1988); the lower-probability conjunction of conditions or events constitutes a more plausible scenario, as a whole, than does one or a subset of them. Dawes illustrates the idea by contrasting the two following event sequences: (a) an alcoholic tennis star wins a major tournament 8 months after beginning to drink a fifth a day, and (b) an alcoholic tennis star begins drinking a fifth a day, joins AA a month later, quits drinking, and wins a major tournament 8 months following the beginning of his drinking bout. The second sequence may be seen as more likely than the first, even though it is implicitly encompassed by the first, because it provides a plausible account of how the tournament win could occur despite the drinking problem. Scenario thinking, Dawes suggests, has the effect of overestimating the probabilities of scenarios that come readily to mind and underestimating the probabilities of those that do not.

Evidence of the conjunction fallacy has been obtained in a number of studies using similar types of problems. Another example from Kahneman and Tversky (1982a) that has been widely cited in the literature involves the following description: "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations" (p. 126). Statistically naive participants were asked to judge the likelihood that specific claims were true of Linda: for example, Linda is a teacher in an elementary school; Linda works in a bookstore and takes yoga classes; ... Among the claims were the following two: (a) Linda is a bank teller, and (b) Linda is a bank teller who is active in the feminist movement. A majority of the participants judged the second of these claims to be more probable than the first. Even graduate students who have had training in statistics often considered (b) to be the more probable. Again it is clear that, at least from a purely statistical point of view, this is the wrong choice because it violates the rule that the probability of the conjunction of two events cannot be greater than the probability of the less probable of the constituent events.

Is the Conjunction Fallacy a Fallacy?

Although the strength of the conjunction fallacy has been shown to vary considerably with the specifics of the task (Morier & Borgida, 1984; J. F. Yates & Carlson, 1986), it is sufficiently common that the question naturally arises as to whether it may have some functional basis. Is there a point of view from which the selection of Alternative (b) in the foregoing example might be considered reasonable? Note that the information provided about Linda in this problem is totally irrelevant to the selection of the more probable of the two statements. Statement (a) is more probable (more accurately, not less probable) than Statement (b) by virtue of the conjunction rule, independently of any information provided in Linda's description. However, it seems not unreasonable of participants in psychological experiments to assume that the information that is being given to them in problem-solving situations is relevant to the problem they are asked to solve. In this case, most of the information provided about Linda is, by intention, the type of information that would be useful for distinguishing activists in the feminist movement from nonactivists but not for distinguishing bank tellers from non–bank tellers. One might say the experimental approach is to lead the participant down a garden path. Perhaps some people select (b) on the assumption that the experimenter would not provide them with irrelevant or misleading information.

It is easy to imagine being given the kind of information about Linda in this example in a nonlaboratory situation in which one was being asked to judge the probability that she was active in the feminist movement, whereas it is much less easy to imagine being given this information if one were being asked to judge whether she was a bank teller. If one were being asked to decide whether she was a bank teller, one would expect to be given information that was more relevant to that decision. The fact that people erroneously select Alternative (b) when presented with this problem may be less than compelling evidence that they are poor statistical reasoners; it may be that they interpret the question according to certain rules of language usage that work in everyday conversation if not always in the psychology laboratory. If the point were to determine whether people understand and can apply the conjunction rule one might ask them to judge which of the two statements, (a) or (b), is the more probable without providing the garden-path information about Linda or indeed any information about her at all.

Kahneman and Tversky (1982a) make essentially this point in one of their discussions of this and similar findings. They cite the "cooperativeness principle" by which the listener in a conversation is entitled to assume that the speaker is trying to be informative (H. H. Clark & E. V. Clark, 1977; Grice, 1975; Searle, 1975). They note that people are likely to assume that the principles that apply to everyday conversations apply also in the experimental laboratory, which is to say they are likely to assume that the information provided them by the experimenter in problem-solving situations is relevant to the task that is to be performed; if it were not, according to the cooperative principle, the experimenter would not provide it. Macdonald (1986; Macdonald & Gilhooly, 1990) makes a similar argument, and raises the question whether, given the conventions of everyday conversation, participants in these experiments are wrong to see the phrasing of questions as providing information. Other investigators have also argued that what have sometimes been taken as evidence of illogicality in dealing with Linda-like problems may have more to do with linguistic conventions (Dulany & Hilton, 1991; Politzer & Noveck, 1991; Slugoski & A. E. Wilson, 1998).

The idea that people take the phrasing of questions as informative gets strong support from the fact that participants in Tversky and Kahneman's (1983) experiment considered it more likely than not that Linda was a feminist bank teller. As Macdonald and Gilhooly (1990) argue, the probability that a randomly chosen woman would be a feminist bank teller is surely very small. One explanation of the relatively high probabilities that the participants assigned to Linda's being a feminist teller, they contend, is that the question itself suggests that there may be reason to believe that she is.

L. J. Cohen (1982), among others, takes the position that the performance of people in these situations can be seen as reasonable behavior, depending on what we assume people understand their task to be. Perhaps they take their task to be to assess the believability of a story about a person cast in terms of the causes or effects of the collections of features constituting their profiles, rather than as that of estimating the relative frequencies of different kinds of people:

> One has to take into account not just the meanings of sentences in which instructions to subjects are formulated but also the implications of uttering them. So

> when asked for the probability of a particular single event subjects may well infer that what is wanted is an estimate of the believability of that event's occurrence as an apparently isolated event, which could well be lower than the believability of the occurrence of a particular causal sequence containing the event. If, on the other hand, subjects are asked specifically for the probability of the single event's occurrence *whether* in isolation *or* within the particular sequence, it would be very surprising indeed if they then went on to declare the particular sequence's occurrence to be even more probable. (p. 264)

L. J. Cohen (1982) argues that with respect to questions of conservatism in the use of data, the gambler's fallacy, and the conjunction fallacy, investigators have chosen to impute a mathematical fallacy to their subjects rather than to impute to them an element of doubt as to whether the experimental situations in which their performance was studied constituted perfect games of chance. He takes the strong position that "no protocols could confirm the view that laymen are inherently prone to overestimate the probabilities of conjunctions or underestimate those of disjunctions, since it is quite clear that when the task is unambiguously presented lay subjects are capable of responding in accordance with correct mathematical principles in regard to the Pascalian probability both of a conjunction (Beach, 1966) and of a disjunction (Beach & [C. R.] Peterson, 1966)" (p. 265).

Another possible account of the conjunction fallacy with problems like the Linda-bank teller one, which I will refer to as the "filling-in" hypothesis, might invoke a convention of language interpretation along the following lines. Given the two statements, (a) "Linda is a bank teller," and (b) "Linda is a bank teller who is active in the feminist movement," a reader or listener might interpret (a) as "Linda is a bank teller *who is not active in the feminist movement."* I am not arguing that one should interpret the statement this way, but it does not seem to require much of a stretch of the conventions of language use to imagine how one might do so. It would be unusual in an everyday situation to contrast the likelihood that Linda is a bank teller who is active in the feminist movement with the likelihood that she is a bank teller; it would not seem strange, however, to contrast the likelihood that she is a bank teller who is active in the feminist movement with the likelihood that she is a bank teller who is not active in this movement. If one puts this interpretation on (a), then judging the probability of (b) to be greater than the probability of (a) is not an example of the conjunction fallacy, because, given this interpretation, the set referenced by (b) is not a subset of that referenced by (a); the two sets are disjoint.

This possibility gets some indirect support from a comment by Dawes (1988) in the context of a discussion of one of Tversky and Kahneman's (1983) studies. Kahneman and Tversky had given medical internists the following problem:

> A fifty-five-year-old woman had a pulmonary embolism (blood clot in the lung). How likely is it that she also experiences:
>
> - dyspnea [shortness of breath] and hemiparesis [calf pain]
> - pleuritic chest pain
> - syncope [fainting] and tachycardia [accelerated heart beat]
> - hemiparesis
> - hemoptysis [coughing blood]. (Dawes, 1988, p. 130)

Dawes (1988) notes that "on the average, 91% of the 32 internists questioned believed that the combination of a probable symptom (in this case dyspnea) and the improbable one (in this case hemiparesis) was more likely than the improbable symptom alone" (p. 130). What I want to call attention to is the ambiguity of this statement. "The improbable symptom alone" could be taken to refer to the improbable symptom (hemiparesis), whether or not accompanied by the more probable symptom, or it could be taken to mean the improbable symptom *by itself,* which is the improbable symptom *in the absence of* the more probable one. These two interpretations are quite different and a conjunction fallacy is implicated only in the case of the former one, because, given the latter interpretation, the two sets—(a) dyspnea and hemiparesis and (b) hemiparesis *alone*—are disjoint. The ambiguity of the terminology in this description of the outcome of an experiment lends credence to the possibility that participants in the experiment could have been confused by the same type of ambiguity.

Appeal either to the cooperativeness principle or to the filling-in hypothesis helps us see how people could be led to overlook the simple set-subset relationships involved in the situations described previously, but neither seems to apply as readily to another result obtained by Tversky and Kahneman. In this case, one group of participants was asked to estimate the frequency of seven-letter words with "n" in the next-to-last position, and a second group was asked to estimate the frequency of seven-letter words ending in "ing." Inasmuch as the first set includes the second it cannot be the smaller of the two. Nevertheless, estimates of the "ing" set were reliably larger than estimates of the "n_" set. One might object to treating this result as an example of the conjunction fallacy, inasmuch as the same participants did not make both estimates. On the other hand, if the estimates the two groups produced are taken as indications of what people generally believe about the sizes of these two sets, we must conclude that the subset is generally believed to be larger than the more inclusive set from which it is drawn.

The final example relates to the conjunction fallacy in an unusual way and was also reported by Tversky and Kahneman (1971). In this case, research psychologists found a single experimental result that strongly supported an hypothesis to be more compelling evidence than the same result combined with another that provided weak, but still positive, support for the hypothesis. It appears that the weight of the strong result was diminished by being coupled with the weak one, even though the conjunction of strong and weak positive results is less likely to occur by chance than is a strong positive result alone.

Gigerenzer (1991b, 1993, 1994) argues that the "conjunction fallacy," as commonly described, is not really a fallacy, at least in the eyes of a probability theorist of the frequentist school. The argument is similar to that used to discount the overconfidence effect (see chap. 8). Frequentists, Gigerenzer contends, do not recognize the meaningfulness of the application of probability to single events, so the question of the probability that a person fitting a given description is a member of a specified profession has no answer: "What is called the 'conjunction fallacy' is a violation of *some* subjective theories of probability, including Bayesian theory. It is not, however, a violation of the major view of probability, the frequentist conception" (1991b, p. 92). "From the frequency point of view, the laws of probability are mute on the Linda problem, and what has been called a conjunction fallacy *is not* an error in probabilistic reasoning—probability theory simply doesn't apply to such cases. Seen from the Bayesian point of view, the conjunction fallacy *is* an error" (1993, p. 293).

In any case, it appears that when problems like those described earlier are phrased in frequentistic terms—for example, how many out of 100 people fitting Linda's description are (a) bank tellers (b) bank tellers and active in the feminist movement?—the conjunction fallacy is much less likely to occur (Fiedler, 1988; Hertwig & Gigerenzer, 1994). Gigerenzer (1994) argues that, this being the case, "instances of the 'conjunction fallacy' cannot be properly called reasoning errors in the sense of violations of the laws of probability. The conceptual distinction between single-event and frequency *representations* suffices to make this allegedly stable cognitive illusion largely disappear" (p. 144). And, Gigerenzer notes, the conjunction fallacy is not the only cognitive illusion to which this argument applies.

CONDITIONAL PROBABILITIES

The probability of A conditional on B, often represented as $p(A \mid B)$, is the probability that A is true, given that B is true. According to the probability calculus $p(A \mid B) = p(A \,\&\, B)/p(B)$, which is to say the probability that both A and B occur, divided by the probability that B occurs. If A almost always occurs when B does, the conditional probability is high; if B more often than not occurs without A, the conditional probability is low.

Difficulties in Conditionalizing

Conditional probability, though simple in concept, proves difficult sometimes to apply in specific situations. Consider, for example, the following problem from Falk (1979, 1983). An urn that contains four balls—two white and two black—is shaken vigorously and two balls are drawn from it, blindly and without replacement. Falk found that when students were asked what the probability is that the second ball is white, given that the first one is white, $p(W_2 \mid W_1)$, they tended to get the correct answer, 1/3. However, when asked what the probability that the first ball to be drawn was white, given that the second one is white, $p(W_1 \mid W_2)$, they had difficulty in seeing that the same answer is correct. To see that the two situations are comparable, consider the following two problems: What is the probability that the third ball drawn is white, given that the first and second ones were white, $p(W_3 \mid W_1\&W_2)$? and What is the probability that the first ball drawn was white, given that the second and third ones are white, $p(W_1 \mid W_2\&W_3)$?

The determination of conditional probabilities in practical situations is often very difficult. In part, this is because, as Edwards et al. (1963) argue, all probabilities can be seen as conditional probabilities. Conventionally Bayesians distinguish between the prior probability of a hypothesis, $p(H)$, and the posterior probability of that hypothesis after the receipt of some data that has relevance for it, $p(H \mid D)$, and treat only the latter as conditional. Edwards et al. point out that the prior probability really is conditional as well as the posterior: "Thus, $p(H)$ is the probability of the hypothesis H for you conditional on all you know, or knew, about H prior to learning D; and $p(H \mid \mathrm{D})$ is the probability of H conditional on that same background knowledge together with D" (p. 199).

Essentially the same point can be made with respect to any probability, at least if one equates probability with degree of belief. Rozeboom (1997) puts it this way: "According to statistical theory, probabilities in a mathematically structured system thereof are always conditional on one or another configuration *P* of population-defining properties variously described as preconditions, background constraints, local constancies, or other phrases similarly connotating limitation" (p. 386). Usually, the many conditions that underlie any particular statement of probability are not made explicit—it is not clear that all of them could be—and in many cases perhaps they are not recognized as conditions. For example, the statement that the probability of obtaining head on the next toss of a coin is .5 is conditional on the coin being fair, the laws of physics not changing before the toss, the tosser not being adept at influencing the outcome, there being no magic or telekinesis at work, and so on. Most such conditions, we would say, go without saying. Often, however, the failure to recognize a material conditionality of a probability and to factor it into a computation can lead to erroneous results.

The importance of correctly identifying conditional probabilities is readily seen in the context of diagnostic decision making. In medicine, for example, it is essential that a clear distinction be made between the probability of a specific symptom or test result being observed conditional on the presence of a particular disease and the probability of a particular disease being present conditional on the observation of a specific symptom or test result. Treating both variables as binary, Table 9.2. shows the possible combinations of test result and disease state.

The probability of a positive test result conditional on the presence of the disease, $p(T+ \mid D+)$, is known as the *sensitivity* of the test; the probability of the presence of the disease conditional on a positive test result, $p(D+ \mid T+)$, is known as the *predictive value of the positive test,* and it is the latter that the physician usually wants to know (B. K. Holland, 1998). If the table entries are frequencies or joint probabilities,

$$p(T+|D+)=\frac{T+D+}{(T+D+)+(T-D+)},$$

and

$$p(D+|T+)=\frac{T+D+}{(T+D+)+(T+D-)}$$

It is easy to see that these probabilities can be greatly different and that confusing them could have unfortunate consequences. Other conditional probabilities of interest that can be computed from the table, $p(T- \mid D-)$ and $p(D- \mid T-)$, represent, respectively, the *specificity* of the test, which is the probability that the test is negative conditional on the disease being absent, and the *predictive value of a negative test,* which is the probability that the disease is absent conditional on the test being negative. B. K. Holland (1998) discusses the relation-

TABLE 9.2
The Possible Combinations of Test Result (*T*) and Disease State (*D*), Treating Each Variable as Binary

	Disease	
Test Result	*Present*	*Absent*
Positive	T + D +	T + D −
Negative	T − D +	T − D −

ships among sensitivity, specificity, and the predictive values of positive and negative tests. An understanding of these relationships is critical to appropriate application of the results of clinical tests in medicine.

We should note too that, as indicated in chapter 4, the likelihood ratio is the ratio of $p(T+ \mid D+)$ to $p(T+ \mid D-)$ (although in the notation of chap. 4, D was used to represent data whereas here it is used to represent disease, which is equivalent to H in the earlier notation) and the diagnosticity of the test would be indicated by the degree to which that ratio differs from 1.

Transposed Conditional Probabilities

A distinction between $p(A \mid B)$ and $p(B \mid A)$ is one that many people fail to make (Bar-Hillel, 1974; Kahneman & Tversky, 1973). The tendency to see these two conditional probabilities as equivalent, which Diaconis and Freedman (1981) call the *fallacy of the transposed conditional,* bears some resemblance to the widely noted *premise conversion error* in syllogistic logic (Henle, 1962). Various explanations of the failure to distinguish between them have been proposed. One suggestion is that they are special cases of a general tendency to see relationships as symmetrical (Tsal, 1977). Baron (1988) raises the possibility that some of the apparently irrational behavior of people on Bayesian reasoning tasks stems from a confusion of $P(A \mid B)$ with $P(B \mid A)$.

The possibility of transposition in reasoning about conditional probabilities has been noted by several writers (Bar-Hillel, 1984; Wyer, 1977; Wyer & Srull, 1989). It gains credence from experimentation by Sherman, McMullen, and Gavanski (1992), who asked people to make conditional frequency estimates of the sort, "Of 100 randomly chosen men, how many prefer blue rather than brown?" and "Of 100 randomly chosen people who prefer blue rather than brown, how many are men?" People were more likely to give evidence of a transposition error in the latter case than in the former. Sherman et al. argue that people have ready access to an appropriate sample space from which they can mentally draw a useful sample when the conditioning event is a natural category (men) but not when it is an unnatural category (people who prefer blue). When the conditioning event is of the latter type, they are likely to invert the problem and base their estimate on the use of the more natural category.

Lack of precision in the reporting of conditional probabilities is often the basis of ambiguous claims about the effectiveness of diagnostic tests. As noted in Chapter 8, the report of an overall accuracy level can mean many different things, and the statistical results of testing can be very sensitive to the way people to be tested are sampled (e.g, whether they self-select because they consider themselves at risk), and this is especially so in the case of diseases that have unusually low incidence in the population. Accuracy figures should be reported as conditional probabilities, and the basis of selection for testing should always be made explicit, if confusion is to be minimized.

SUMMARY

The ability to detect covariation is widely recognized as an important one, but the extent to which people manifest this ability has been questioned by many researchers. Some have contended that in the absence of specific training most people lack the ability to think in correlational terms; others have argued that many of the experimental results that support this conclusion have been obtained in laboratory situations that are not adequately representative of the kinds of real-life problems for which the detection of covariation would be advantageous.

A common finding of laboratory studies of covariation perception has been that, when presented with contingency data of the kind that can be represented in a two-by-two table showing the four possible combinations of the presence or absence of two variables of interest, people tend to focus too much attention on, or give too much weight to, the cases in which both variables occur— the "plus-plus" cell of the contingency table—and tend to neglect or discount those in which one or both variables are absent. This bias of focus is said to lead often to an overestimation of the strength of the relationship between the variables of interest.

Prominent among the variables that researchers have identified as determinants of the perceived degree of covariation are expectations and preexisting beliefs about cause–effect relationships. Given the same statistical data, perceived degree of covariation tends to be greater when the covariation is consistent with expectations and especially when people believe there to be a cause–effect relationship between the variables involved.

Considerable research has focused on the phenomenon of "illusory correlation," which is a perceived correlation between variables that are not correlated in fact, or a perceived correlation that is materially higher than the actual one. Much of this work has centered on perceived correlations between or among personality traits and patterns of behavior or on perceived consistency of behavior over time.

Some studies have revealed common tendencies to overestimate the probabilities of conjunctions of events and to underestimate the disjunctions of events, relative to estimates of the probabilities of the component events. Such tendencies can help account for what appears to be a common predilection to undue optimism in predicting the time or effort that will be needed to complete a specific task.

Possibly the most extensively studied phenomenon involving the perception of covariation is what is widely referred to as the "conjunction fallacy." Although the probability of the conjunction of two events cannot be greater than the probability of the more probable of the two, it appears that in many contexts people are likely to violate this principle when judging the likelihood of conjunctions. The phenomenon has stimulated much research and theorizing. Debate continues as to whether it represents a genuine reasoning fallacy or simply a reflection of the recognition of certain linguistic conventions.

Dealing with conditional probabilities appears to be problematic for many people. A common confusion is between the probability of *A* conditional on *B* and the probability of *B* conditional on *A*. This type of confusion is analogous to a confusion often made in logic in which "If *A* then *B*" is sometimes seen as equivalent to "If *B* then *A*." The confusion can be aggravated by the ambiguous reporting of conditional probabilities, as, for example, when a diagnostic test is said to have a specified level of accuracy, without clarification of precisely what that means.

CHAPTER

10

Choice Under Uncertainty

Simply characterized, rational decision making consists in one's choosing the best member from the set of available alternatives.... The central question for a theory of rationality concerns, of course, how someone is rationally to assess the available actions in a decision-situation.

—*Moser (1990, pp. 2, 3)*

A satisficer is concerned with doing well enough, while an optimizer is concerned with doing the best it can.

—*Goodie, Ortmann, J. N. Davis, Bullock, and Werner (1999, p. 327)*

PRELIMINARY DISTINCTIONS

Choice or decision-making situations—"choice" and "decision making" are used synonymously here—can be classified in several ways. One basic distinction that has been made is between "yes–no" and "forced-choice" situations. The latter type is usually taken to be the prototypical model of decision making; the problem here is to choose one from among several alternative courses of action. However, the "yes–no" case is not uncommon in operational contexts in which the decision maker has only a single action possibility and the problem is to decide whether to not to take it (Mintzberg, 1975; Peters, 1979). Of course, the "yes–no" situation can be seen as a special case of "forced-choice," namely that in which the choice is forced between taking the one possible action or declining to do so.

Outcome Certainty or Uncertainty

Another fundamental distinction relates to the degree to which the outcome of a choice is known in advance. At one extreme is the case in which the outcome is known with certainty: The choice is between *A* and *B,* and what one chooses is what one gets. The only problem here is deciding which of the alternatives one prefers. In contrast is the case of risky decision making or decision making under uncertainty, in which the outcome of a choice is known only probabilistically.

Making choices under uncertainty is difficult because one must take into account not only what one's preferences are but the fact that they are not assured of being realized in any case. How to measure and deal with uncertainty in decision-making situations has been, and continues to be, a major focus of decision-making research. If a situation is uncertain, one is effectively using more information if one takes account of that uncertainty than if one ignores it, and in some instances, the expected detrimental effect of ignoring uncertainty can be substantial (M. G. Morgan & Henrion, 1990).

The case of incompletely known outcomes can be subdivided into two subcategories, reflecting different degrees of knowledge about the decision context. On the one hand are those instances in which the decision maker knows the available action alternatives and their outcomes, given hypothesized states of the world, and can assign probabilities to the states of the world, thus permitting probabilistic inferences regarding consequences of action selections—these instances represent what is called decision making under *risk* of *measurable uncertainty.* In contrast are those cases in which the decision maker does not know the possible states of the world even probabilistically, and so cannot infer probabilistically the consequences of action selections. These cases are referred to as decision making under *unmeasurable uncertainty* (Knight, 1921; Moser, 1990). Sometimes *uncertainty* (unqualified) is used to refer to *unmeasurable uncertainty* in contrast to *risk.* More often, I think, it is used more generically to connote both kinds of uncertainty, thus subsuming risk. Most of the research that has been done on decision making with incompletely known outcomes has involved decision making under *risk* or *measurable uncertainty,* according to this distinction, but the distinction has not been sharply maintained in the literature and research has often been described simply as involving decision making under *uncertainty* without qualification as to the type of uncertainty involved. The distinction is not emphasized in this book.

Utility and Value

> It is apparent that one franc has much greater value for him who possesses only a hundred than for a millionaire. (Laplace, 1814/1951, p. 22)

Daniel Bernoulli is generally credited with being the first to realize that the desirableness of money or of anything with cash value does not necessarily increase linearly with the amount involved. The intuitive compellingness of this insight is seen in the observation that the amount of pleasure one would derive from an unexpected windfall of a fixed amount of money is likely to depend on how much money one has already; a gift of $1,000 is likely to be a much greater cause of celebration for a person who is broke than for one who already has $1 million salted away.

Bernoulli questioned the reasonableness of considering irrational the behavior of a pauper who, finding a lottery ticket known to be worth, with equal probability, either 20,000 ducats or nothing, decides to sell it for 9,000 ducats, 1,000 ducats less than its expected value. It was this type of situation that led Bernoulli to make a distinction between cash value and utility (or moral worth) and to speculate that the utility of an individual's wealth increased as the logarithm of its cash value. Laplace made much of this distinction between, in his terms, *fortune physique* and *fortune morale*. A preference for certain outcomes over uncertain outcomes with greater *expected* value has been found many times since Bernoulli made this observation, and not only among paupers (Allais 1979/1990; Kahneman & Tversky, 1979b).

P. L. Bernstein (1996) refers to the idea that utility is inversely related to the worth of what one already possesses as one of the great intellectual leaps in the history of ideas. The hypothesized decelerating rate of increase in utility with increasing cash value has the interesting consequence that negative utility of a loss of a given value will be greater (in absolute terms) than the positive utility of a gain of the same value. Bernstein puts it this way: "The logical consequence of Bernoulli's insight leads to a new and powerful intuition about taking risk. If the satisfaction to be derived from each successive increase in wealth is smaller than the satisfaction derived from the previous increase in wealth, then the *dis*utility caused by a loss will always exceed the positive utility provided by a gain of equal size" (p. 112).

MODELS OF DECISION MAKING

There have been many theoretical treatments of decision making (Howard, 1966; Kahneman & Tversky, 1979b; Keeney & Raiffa, 1976; Raiffa, 1968; Raiffa & Schlaifer, 1961; L. J. Savage, 1954; von Neumann & Morgenstern, 1953; von Winterfeldt & Edwards, 1986; Watson & Buede, 1987). (Poundstone

[1992] refers to von Neumann and Morgensterns's *Theory of Games and Economic Behavior* as "one of the most influential and least-read books of the twentieth century" [p. 41].) And numerous models of human choice behavior have been developed. Hogarth (1987) distinguishes among seven such models, three of which (a linear model, an additive-difference model, and an ideal-point model) are considered compensatory models because they recognize the possibility of trade-offs between choice dimensions, and four of which (a conjunctive model, a disjunctive model, a lexical-graphic model, and an elimination-by-aspects model) are considered noncompensatory because they do not recognize such trade-offs. Deciding which of these or other models is most descriptive of human choice behavior is difficult because some of them make the same predictions about performance while assuming different underlying processes.

The problem of trade-offs arises when in order to increase the attractiveness of a decision alternative with respect to some property or dimension, one must decrease its attractiveness with respect to another. The Federal Drug Administration's (FDA) responsibility for deciding how long to test a new drug before approving its clinical use illustrates the problem: The longer a drug is tested, the more accurately its effects, including problemsome side effects and effects arising from interactions with other drugs, can be determined, but the longer it is tested the longer it remains unavailable to people who could benefit from its use. Under pressure to accelerate its approval process, the FDA has introduced regulations that are intended to make drugs available relatively quickly at an acceptable risk of approving some drugs that later prove to be ineffective or deleterious (Kessler & Feiden, 1995). What constitute fast enough and acceptable levels of risk are matters of continuing debate.

Berl, Lewis, and Morrison (1976) attempted to compare the adequacy of several models of riskless choice as applied to the problem of college selection by prospective students. The models they considered were additive weighting (Edwards & Tversky, 1967), satisficing (Simon, 1955), and lexicography (Coombs, 1964; Luce, 1956; Tversky, 1969). The additive weighting model produced somewhat more accurate predictions of choices than did the alternative models. However, Berl, Lewis and Morrison, in keeping with several other investigators, preferred to view the additive weighting model as a black box that predicts choices rather than as a description of how the choices are made. They cite evidence that when people believe themselves to be using an additive weighting model, they tend to overestimate the number of factors used (Shepard, 1964) and to give estimates of weights on the factors that differ from those that best reproduce their decisions (Dawes & Corrigan, 1974; Slovic & MacPhillamy, 1974).

Much work has been done on the development of normative models of choice and decision making. Some of that work has already been mentioned in this book, especially in the context of the discussion of Bayesian reasoning. Numerous tutorials, reviews, and anthologies are available to the interested reader, including Luce and Raiffa (1957), Howard (1968), North (1968), Raiffa (1968), R. V. Brown, Kahr, and C. R. Peterson (1974), Dawes and Corrigan (1974), M. F. Kaplan and Schwartz (1977), Slovic et al. (1977), Hammond, McClelland, and Mumpower (1980), Einhorn and Hogarth (1981), Arkes and Hammond (1986), von Winterfeldt and Edwards (1986), Dawes (1988), Dowie and Elstein (1988), Berger (1997), and Hammond (2000).

Normative approaches to decision making under risk often involve conceptualizing the decision situation in terms of a set of states of the world that could conceivably hold and a set of decision alternatives, assigning probabilities to the possible states of the world, assigning a value (utility; gain-loss) to each of the possible combinations of states of the world and decision alternatives, and then selecting an alternative according to some goal. The situation may be represented abstractly by a matrix, with one dimension (say the rows) indicating the possible, or hypothesized, *states of the world,* and the other (the columns) indicating the *action alternatives* or options available to the decision maker. Each cell entry in the matrix indicates the relative desirabilty—the *utility*—to the decision maker of the outcome that would result if the associated action alternative were selected and the associated state of the world pertained. In the following representation, for example, U_{ij} represents the desirability to the decision maker of the outcome resulting from the choice of Action Alternative *j* if the real state of the world is as indicated by Hypothesis *i*.

	Action Alternatives						
Hypothesized States	A_1	A_2	A_3	...	A_j	...	A_n
H_1	U_{11}	U_{12}	U_{13}	...	U_{1j}	...	U_{1n}
H_2	U_{21}	U_{22}	U_{23}	...	U_{2j}	...	U_{2n}
H_3	U_{31}	U_{32}	U_{33}	...	U_{3j}	...	U_{3n}
.							
H_i	U_{i1}	U_{i2}	U_{i3}	...	U_{ij}	...	U_{in}
.							
H_m	U_{m1}	U_{m2}	U_{m3}	...	U_{mj}	...	U_{mn}

The hypothesized states of the world are assumed to be an exhaustive and mutually exclusive set, which is to say that it is assumed that one and only one of them pertains. Associated with each hypothesized state, H_i, is a probability, $p(H_i)$ and the assumption of an exhaustive and mutually exclusive set means that these probabilities sum to 1, $\sum_{c=1}^{n} p(H_i) = 1$. The set of action alternatives also is assumed to include all there are and the decision maker can choose only one of them. The values of the individual cells can be expressed in any consistent units to reflect the relative desirability to the decision maker of the possible outcomes.

Many normative models of decision making begin with this representation and provide algorithmic rules for selecting an action alternative so as to realize some goal, which could be to maximize one's expected utility, to minimize one's maximun possible loss, or to accomplish something else. The goal one selects is likely to depend on such factors as the stakes, one's willingness to accept risk, whether one expects to make many decisions or only one, and one's general outlook. It is clear from the representation, however, that major challenges to the decision maker are those of identifying what the action alternatives and possible states of the world are, assigning probabilities to the latter, and quantifying preferences (utilities) for the various possible outcomes. The preceding representation of decision making is often seen in discussions of Bayesian reasoning; what Bayes's rule (see chap. 4) provides is a way of using newly acquired evidence to modify an existing distribution of probabilities over the hypothesized states of the world, or to update, in other words, the probabilities associated with the various hypotheses in the light of new data that relate to them.

MAXIMIZATION OF EXPECTED UTILITY

> Once the theory of probability has been taken for granted, the principle of maximizing the expected utility per unit time [or rather its integral over the future, with a discounting factor decreasing with time, depending on life expectancy tables] is the only fundamental principle of rational behavior. (Good, 1983a, p. 9)

Maximization of expected utility is one of the goals that can be used to guide choices, given the analytic approach described previously. Some theorists, like Good (1983a), have taken the strong position that the principle of maximizing expected utility, in one or another form, is the only defensible guide for rational action. Expected utility theory has been widely interpreted not only as a normative theory of rationality, but also as a descriptive theory of how people actually behave or at least wish to behave (Arrow, 1971; Friedman, 1953; Friedman & L. J. Savage, 1948). On the other hand, the use of subjective utility in models of choice is not without critics: "Rational choice models make heavy use of both subjective utilities and subjective probabilities, as well as of the simplistic hypothesis that selfishness is the only motivation of human behavior. Not surprisingly, none of these models fits the fact.

Hence, although at first sight they look scientific, as a matter of fact they are pseudoscientific" (Bunge, 1996, p. 104).

Despite the usefulness of the principle of maximization of expected utility and its popularity among economists, it has problems, some of which have been known for a long time. There seems to be a growing consensus among investigators of decision-making behavior that the classical view of decision making, which promotes maximization of expected utility as a goal, and analysis of the situation into possible world-states and action alternatives, is not descriptive of the way most decisions are actually made, including decisions with significant stakes that are made under conditions that would permit an analytic approach. As Beach, B. Smith, Lundell, and Mitchell (1988) put it, "Whatever else may have been learned in thirty-five years of behavioral decision research, the primary lesson must be that expected utility and its theoretical accoutrements provide an unsatisfactory description of decision-making at the individual level" (p. 17). If behavior that maximizes expected utility is taken as the norm for rationality, then there seems no alternative to the conclusion that people behave irrationally more often than not, which is to say that, as a descriptive model of human behavior, maximization of expected utility misses the mark (Miller & Starr, 1967). In fact, the adequacy of expected utility theory as a normative model of rationality has also been the subject of some debate (Beach et al., 1988; Krantz, 1991; Sahlin, 1987; Shafer, 1986).

Keynes (1921/1956), for example, rejects the idea that "in order to obtain ... a measure of what ought to be our preference in regard to various alternative courses of action, we must sum for each course of action a series of terms made up of amounts of good which may attach to each of its possible consequences, each multiplied by its appropriate probability" (p. 1363). His rejection of this position is based in part on the grounds that not all the assumptions that underlie it are justified, and in part also on the grounds that not only probabilities but the weight of evidence on which they are based should be taken into account:

> If, therefore, the question of right action is under all circumstances a determinate problem, it must be in virtue of an intuitive judgment directed to the situation as a whole, and not in virtue of an arithmetical deduction derived from a series of separate judgments directed to the individual alternatives each treated in isolation. We must accept the conclusion that, if one good is greater than another, but the probability of attaining the first less than that of attaining the second, the question of which it is our duty to pursue may be indeterminate, unless we suppose it to be within our power to make direct quantitative judgments of probability and goodness jointly. (p. 1364)

Keynes questions the assumption that goods and probabilities can be combined multiplicatively in a straightforward way:

> Is it certain that a larger good, which is extremely improbable, is precisely equivalent ethically to a smaller good which is proportionately more probable? We may doubt whether the moral value of speculative and cautious action respectively can be weighed against one another in a simple arithmetic way, just as we have already doubted whether a good whose probability can only be determined on a slight basis of evidence can be compared by means merely of the magnitude of this probability with another good whose likelihood is based on completer knowledge. (p. 1366)

The difficulty is compounded by the fact that what is to be maximized according to the maximization-of-expected-utility model is a number that is the product of two terms, one an indicant of uncertainty and the other an indicant of worth, each of which can be treated as either an objective or a subjective variable. One can use as an indicant of uncertainty either an objective measure of probability, say a relative-frequency statistic, when available, or a subjective measure, like the decision maker's personal belief regarding what is or is not likely to be the case. Similarly, as an indicant of worth, one can use something as objective as monetary value, or something more subjective—utility—that, like subjective probability, can vary from person to person in the same situation. Moreover, models of decision making can be constructed by combining the objective and subjective forms of each variable in any of the four possible ways (Coombs, Bezembinder, & Goode, 1967). Although things would be greatly simplified if models could be based on the use of objective measures for both variables, there has long been a strong consensus among decision theorists that such models are neither appropriate as prescriptions for behavior nor descriptive of how people actually behave.

The Quantification of Expectation and Utility

One major difficulty is that of quantifying expectation. When an event is to be determined by some random process that is understood by the decision maker (e.g., toss of a coin, roll of a die) or when relative-frequency data are available, there is a way to link expectations to objective measures of probability, but things become more vague and abstract when the events of interest are of the type that have never occurred in the past (nuclear war between superpowers, exhaustion of fossil fuel reserves), so they have no relative frequency of occurrence, or that cannot occur many times in the future so that even imagined relative frequencies make dubious sense. It is not clear what expectation means in such cases and how it should be quantified. It has been argued that many real-life decision situations are unique events to those who face them, and that therefore decision makers are rightfully more interested in the question of possibilities than in that of probabilities; the idea of probability is meaningless, the argument goes, when applied to one-of-a-kind events (Gigerenzer, 1987; Shackle, 1958/1967).

At least as serious as the problem of quantifying expectation is that of determining what people value in specific situations—the problem of measuring utility. It can be very difficult to demonstrate that an attempt to maximize utility is not being made in any particular case, because one can always ascribe utility to some aspect of the situation that has the effect of making the individual's behavior rational by definition. This is similar to the possibility of defining action that is consistent with one's perceived self-interest as action that one elects to take, which makes unselfish action impossible in principle. One can make relationships "vacuously true," to use Buchanan's (1978) phrase, by definition, but in doing so one makes them uninteresting.

Consider again, for example, the results from "probability-matching" experiments in which people often behave in what appears to be an irrational way. In the simplest version of this experiment, the task is to predict on each of a series of independent trials which of two chance events—say a red or a green light—will occur on that trial, and one's earnings for a session depend on how many predictions are correct. As already noted, the strategy that will maximize the expected number of correct predictions in such a situation is to predict *always* the more frequent event; as soon as it becomes clear that the red light, say, occurs more frequently than the green one, one should, from that point on, always predict red. But people typically do not do this. Instead of adopting the strategy that would maximize the number of correct predictions, they tend to predict the more likely event with about the frequency of its occurrence.

On the face of it, probability-matching behavior seems to be irrational, at least from the point of view of maximization of expected utility; we might rationalize it, however, by ascribing a high utility to keeping the situation interesting for the participant. To select the same alternative on every trial might make the situation unacceptably boring to some people, even though this would maximize their expected earnings. Given that the amount of money involved usually is small, the satisfaction of "playing the game" in an interesting way may more than offset the small amount of earnings that one gives up in order to do so. That the playing of the game, or the act of gambling, has no intrinsic utility is typically considered an assumption that is essential to the application of expected utility models to human behavior, but the assumption is questionable in many cases.

Or consider again the individual who buys a ticket in a lottery; to be specific, let us say a $1.00 ticket in a lottery in which there is one chance in 2,000 of winning $1,000. Is such an individual acting irrationally by paying $1.00 for a ticket representing an expected value of $0.50? An affirmative answer gives no value to the act of participating in the gamble. It fails to recognize that one might consider the thrill of playing to be worth a dollar even though one recognizes the chance of winning to be small. From this perspective, what one bought with one's dollar is not simply a 1-in-2,000 chance of winning $1,000, but a bit of excitement and pleasure. Death-risking, daredevil feats can be seen as rational if one recognizes the utility to the risk taker of the thrill itself, not to mention the reinforcement (money, fame, adulation) that success could bring.

As one method for measuring utilities, von Neumann and Morgenstern (1953) proposed the use of a technique that presents the decision maker with an option between a certain situation (e.g., present state of health, present job, known amount of money in the bank) and a gamble, one possible outcome of which is highly desirable (cure from a serious disease, much better job, large gain from financial venture) whereas the other is highly undesirable (death, loss of job, financial ruin). The probability of the undesirable outcome is assumed to be the complement of the probability of the desirable outcome, and one's task is to adjust these probabilities until one is indifferent to the choice between the certain situation and the gamble. The more one likes the certain situation, the higher the probability of the more desirable outcome must be before one will prefer the gamble.

If the utilities of the outcomes of the gamble are known or can be inferred, this procedure provides a means of inferring from them the utility of the certain state. In theory, the number obtained should not depend on the composition of the gamble because the decision maker's positioning of the indifference point will take into consideration the utilities of the possible outcomes. In fact, however, changing the gamble can affect the derived utility (Llewellyn-Thomas et al.,1982/1988; Tversky & Kahneman, 1981).

Moreover, as it happens, people often show a preference, like Bernoulli's pauper, for a guaranteed gain of a specified amount over a gamble with a higher mathematical expectation. For example, given the choice between receiving $800 for sure, and an 85% chance of winning $1,000 coupled with a 15% chance of winning nothing, most people prefer the sure thing over the gamble, even though the expected value of the gamble ($850) is greater than the value of the sure thing. Kahneman and Tversky (1984) refer to the phenomenon as the *certainty effect*. This preference can be seen as rational from a maximization-of-expected-utility perspective by giving some negative utility to risk. (The point that a small sure thing may be preferred to an uncertain outcome with a very large expected value is made by the extreme example of the St. Petersburg paradox, which is discussed in chap. 6.) A similar preference may be seen when one's choice alternatives include no sure thing. When one's choice is between a low-probability alternative with a high value and a high-probability alternative with a lower value, one may well prefer the second option even if the expected value (or utility) is higher for the first one (Lichtenstein & Slovic, 1971, 1973; Tversky, Sattath, & Slovic, 1988).

The preference for the $1.00 lottery ticket that has an expected value of $0.50 is not an example of this generalization, because in this case the value of the dollar in hand exceeds the expected value of the ticket, but it too illustrates the need for the distinction between value and utility; one who believes that people always attempt to maximize expected utility would argue that although the value of one dollar is greater than the expected value of the ticket, the utility of a dollar is less than the expected utility of the gamble in the utility systems of those who would prefer the ticket.

This all seems intuitively reasonable. Most of us probably would agree that neither expected monetary value nor actual monetary value, when it is known, is necessarily an accurate indication of what something is worth to us personally. Moreover, we are not surprised to discover that a wilingness to pay *X* dollars for a certain amount of some good is not compelling evidence that one would be willing to pay 2*X* dollars for twice as much of the same good; if I am traveling without a refrigerator, I may gladly pay the going price for as much ice cream as I can comfortably eat, but I am unlikely to be interested in buying twice as much even at considerably less than twice the price. So the distinction that decision theorists make between monetary value and utility is an easy one to accept intuitively. Unfortunately, whereas monetary value typically is relatively easy to specify, utility is not. In order to see whether an individual's behavior is consistent with an attempt to maximize expected utility, one must know what the individual really values and that may be very difficult to determine, except by inference from one's choice behavior, and that is what utility theory is intended to predict.

One view of the maximize-expected-utility principle is that it is appropriate only for a long-range perspective, which is to say, if one makes decisions so as to maximize expected utility one can be reasonably sure of maximizing actual utility, but only in the long run. People often take a relatively short-range view, however, and attach considerably more importance to what they think will happen soon than to what they anticipate in the future. This is known as discounting the future, and the discount rate sometimes appears to be a rather steep function of time.

The Principle of Invariance

A basic assumption underlying expected utility theory is that people's utilities are constant and that different methods of measuring them will give the same results. In particular, preferences should not depend on the way in which the alternatives are described (assuming accurate and understandable descriptions) or on the way the preferences are elicited (Tversky & Kahneman, 1981). This is known as the principle of invariance, the two aspects of which are referred to as description invariance and procedure invariance.

The finding that neither description invariance nor procedure invariance holds (Fischhoff, Slovic, & Lichtenstein, 1980; Kahneman & Tversky, 1979b) has been seen as strong evidence against the adequacy of expected utility theory as a descriptive account of human choice or decision making. It has also motivated interest in the idea that preferences may sometimes be constructed, rather than only being revealed, as a consequence of the attempt to make them explicit (Slovic, 1995).

Consider a gamble in which one alternative, *A*, offers a high probability of a modest win and a low probability of a considerably larger loss, and the other alternative, *B*, offers a low probability of a large win and a high probability of a small loss. Depending on the amounts involved, people will sometimes express a preference for *A*, if given a chance to make one or the other wager, but will state a higher price for *B* if asked how much they would charge to sell either opportunity to someone else (Lichtenstein & Slovic, 1971, 1973). This type of failure of the principle of invariance, which is referred to as *preference reversal*, has received considerable attention both from psychologists and from economists since it was first reported, and several explanations of it have been offered (W. Goldstein & Einhorn, 1987; Grether & Plott, 1979; Lichtenstein & Slovic, 1971; Loomes & Sugden, 1983; Schkade & E. J. Johnson, 1989; Slovic, 1995; Slovic & Lichtenstein, 1983; Tversky, Slovic, & Kahneman, 1990).

Tversky et al. (1988) suggest that people may approach the task of choosing between pairs of options of the kind that have been used to demonstrate preference reversal in a different way from that in which they approach the task of matching pairs so as to make them equal in value. In the first case, they argue, people are likely to make qualitative comparisons, selecting the alternative that is prefered when one considers primarily the feature or attribute that is deemed most important. The matching task, they suggest, is likely to evoke a more computational approach that attempts to take more than a single attribute into account and to do so in a quantitative way. The two approaches are not guaranteed to yield the same preferences, so the hypothesis is compatible with the fact that reversals occur.

Whatever its explanation, the phenomenon of preference reversal can be used to make people function as money pumps (Berg, Dickhaut, & O'Brien, 1985). It is a problem for any theory that assumes that people have fixed preferences that just need to be revealed, as does traditional expected utility theory. Efforts to discredit or explain away the phenomenon in the interest of protecting the integrity of expected utility theory as a description of human choice have not been very successful (Slovic, 1995).

Utility Maximization and Goals

Optimality can only be determined within a specific frame of reference; what is optimal relative to one value system may be far from optimal relative to another. The individual who values honor above life, for example, will have a different utility function in certain risky decision situations than will one whose values dictate the goal of survival at any cost. What is optimal as judged by a utility function that takes into account only monetary and equally objective factors may be suboptimal when such intangibles as personal satisfaction and peace of mind are added to the equation.

Considering such subjective variables complicates the application of quantitative models to the evaluation of choices or choice behavior, because these variables are difficult to quantify, but to ignore them runs the risk of leaving out of consideration factors that may be among the most important to the decision maker, and consequently of representing the real situation inaccurately. This is a very important point, especially as it relates to the problem of judging the rationality of people's behavior. If A's behavior appears to be irrational to B, in the sense that A is not maximizing expected utility, it may be because B's model of A's utility function is incorrect and, in particular, that B does not understand the importance to A of certain subjective variables that are private matters.

The maximization-of-expected-utility model allows for differences in individuals' utility functions, so one might argue that these concerns do not reflect limitations of the model per se but of the ways in which it is applied. But the ways in which individual differences in utility functions are accommodated can come close to defining utility as whatever it is that one is trying to maximize. Some theorists have argued that even the utility that a given individual assigns to a given decision outcome should be considered a random variable, thus allowing different preferences among the same alternatives at different times (Becker, DeGroot, & Marschak, 1963).

The conception of rationality as optimal behavior assumes one's objectives as given—given that one's goal is *X*, then rational behavior vis-à-vis that goal is defined as *Y*. The conception provides no insights on the question of what constitutes rationality in the selection of top-level goals. Simon (1983/1990) makes this point in observing that the theory of maximization of expected utility finesses completely questions of origins of values or accuracy of facts: "At best, the model tells us how to reason about fact and value premises, it says nothing about where they come from" (p. 195). Simon argues also that application of the theory in the real world would be impossibly difficult for mere mortals without resort to drastic simplifying assumptions: "Human beings have neither the facts nor the consistent structure of values nor the reasoning power at their disposal that would be required ... to apply SEU [subjective expected utility] principles" (p. 197). Efforts to make decision models that are based on the idea of maximation of expected utility more descriptive of human behavior have had the effect of making the models ever more complex. Of course the fact that a complex model can describe or predict human behavior, if it can, does not mean that people actually carry out the same computational steps as does the model.

Closely related to the fact that expected utility theory takes one's goals as given is the assumption that people know what their values are that relate to specific choices even when they have not had occasion to make them explicit. As we have already noted, this is an assumption that not everyone is willing to make. Arrow (1990) puts the matter this way: "The question may be raised how we can possibly know about hypothetical choices if they are not actually made. This is not merely a problem of finding out about somebody else's values; we may not know our own values until put to the crucial test" (p. 340). One might add that the choices that people make in situations simulated in the psychological laboratory are not necessarily reliable indications of what people would do if faced with the same choices in nonlaboratory situations where the stakes are significant and the consequences real. This is not to suggest that participants in experiments deliberately behave differently than they would in real-world situations, but rather that what they do in the laboratory reflects what they *think* they would do outside it, whereas what they would really do could be quite different, and there is no way for them or the experimenter to know for sure, short of observing their behavior in the real-world situations of interest.

A different type of limitation of the principle of maximization of expected utility as the primary standard of rationality has been noted by Gauthier (1986/1990). He points out that one who invariably behaves in such a way as to maximize the expected utility for every choice and is known to do so is likely to be excluded from participating in certain situations that require cooperation:

> The important point in our argument is that one's disposition to choose affects the situations in which one may expect to find oneself. A straightforward maximizer, who is disposed to make maximizing choices, must expect to be excluded from cooperative arrangements that he would find advantageous. A constrained maximizer may expect to be included in such arrangements. She benefits from her disposition, not in the choices she makes, but in her opportunities to choose. (p. 330)

One might argue that the essential difference between the straightforward maximizer and the constrained maximizer is that the latter simply evaluates utility in a broader frame of reference than does the former. But pressed to its limits, this argument would allow the interpretation of all behavior as consistent with an effort to maximize expected utility—broadly conceived—over a lifetime, making the notion uninteresting if not tautological.

The point was made earlier that the goal one selects in any specific decision situation is likely to depend not only on the specifics of the situation but on more abiding factors such as one's willingness to accept risk and one's general outlook on life. Good (1983a) argues the importance of the latter factor by pointing out that the mimimax strategy (the strategy of minimizing one's maximum possible loss, which is an alternative to maximizing expected utility or gain) is an ultraconservative one, reasonable if, and only if, the least favorable initial distribution is reasonable according to one's body of beliefs. The minimax solution assumes, he suggests, that we live in the worst of all possible worlds. Whether or not one accepts this assessment, it is clear that the minimax strategy will produce suboptimal results in many situations. If one used it to guide one's investment decisions, one would do relatively well—less poorly than others—only when almost all investments were doing poorly. When most investments were doing reasonably well, this strategy would leave one far behind the norm.

ALTERNATIVES TO EXPECTED UTILITY THEORY

The view that maximizing expected utility "is the only fundamental principle of rational behavior" (Good, 1983a, p. 9) has been a popular one among economists and decision theorists, but it is not the only one that can be held, and it does not lack critics (Kreps, 1990). Some people find it reasonable to hold that the deliberate choice of an action that is known not to be the best possible choice for oneself is not irrefutable proof of irrationality. Slote (1985, 1986), for example, defends the position that moderation in one's desires—a feeling of being well enough off and a failure to seek to be better off, even when being better off is within one's reach—is not necessarily grounds for concluding that one has lost one's senses. A counter to this position is that when one appears deliberately to choose something that is known, by the chooser, to be less consistent with the chooser's best interests than another option not chosen, the problem is in our understanding of the chooser's utilities; if we understood what the chooser's values really are, we would see that the choice does, in fact, represent an effort to do what is in the chooser's best interest, in the chooser's view. But again, this makes the maximization of expected utility correct by definition, and therefore an uninteresting principle.

Simon's (1955, 1957) proposed alternative to expected utility theory is what he calls the theory of bounded rationality. The bounded rationality that Simon describes is not an optimizing rationality but a gets-by rationality. It is the kind of rationality that yields good-enough results to support the survival, or even the prosperity, of the species, but not to ensure the best possible choice in all situations; and—this is perhaps the most important point—it is manageable by creatures with the cognitive and computational limitations of human beings. This rationality rests on capabilities for focusing attention, for generating alternatives for action, for acquiring facts about the environment, and for drawing inferences from those facts.

Kahneman and Tversky (1979b) have developed a theory of choice behavior, which they call "prospect theory," that is intended to deal with some of the shortcomings of expected utility theory. According to this theory, the choice process is composed of two phases: an early editing phase and a subsequent evaluation phase. During the editing phase the possible decision outcomes, or prospects, may be modified in certain ways so as to increase the ease with which a selection among them can be made. For example, if two competing prospects are identical in some respects and different in others, the ways in which they are identical may be disregarded, or "canceled," for purposes of comparison and selection. When the choice is made it is made not between the prospects per se but between the edited representations of the prospects. Moreover a choice situation is encoded in the decision maker's thinking in terms of what can be gained or lost as a consequence of a decision rather than in terms of the decision maker's final state of wealth or welfare, and, although the idea is not unique to prospect theory, Kahneman and Tversky note that losses tend to loom larger than gains of the same amounts. The editing phase is assumed to be context sensitive, so the same prospect could be edited in different ways when encountered in different contexts.

When the editing phase has been completed the decision maker evaluates the edited prospects and picks the one with the highest value. Value here is assumed to be determined by two scales: a weight, which is somewhat analogous to probability in expected utility theory but—being somewhat less constrained—not equivalent to it, and the subjective value or worth of the prospect. Thus conceived, prospect theory is able, in Kahneman and Tversky's view, to accommodate some of the ways in which human decision making typically violates the axioms of expected utility theory. Prospect theory is qualitatively similar to expected utility theory; its greater descriptive flexibility comes from the use of weights that are not strictly probabilities and the assumption that outcome values are determined by changes rather than by final states.

A strong case that many decisions are dictated by certain intuitive principles, or rules of thumb, that appear to be widely honored has been made by Baron (1998): "Do no harm," "maintain the status quo," "do not go against nature," "be loyal to your group," and so on. Such rules, Baron notes, often work well, but they become problematic when they are elevated to the status of principles that should never be violated. Slavish adherence to the "do-no-harm" principle, for example, would effectively rule out many, if not most, actions that might be taken to accomplish objectives that are universally recognized as desirable, because almost all such actions are likely to have negative effects for someone. A more reasonable principle—and perhaps one that is often intended by advocates of "do no harm"—is "try to keep the negative consequences within acceptable bounds." In this respect these principles are like other rules of thumb, or heuristics; they are useful as long as they are recognized for what they are, but can become bothersome when their limitations are not borne in mind.

INDUCTIVE HEURISTICS

The term *heuristic* is used in two somewhat different ways in the psychological literature. Sometimes it connotes a strategy or rule of thumb that is used to advantage in solving problems for which algorithmic solutions are either not known or impractical. The solutions that are obtained are not assumed to be optimal, but they are can be good enough and they can be effected with limited resources: "Simple heuristics can provide good solutions without a great deal of complex general purpose calculation" (Goodie et al., 1999, p. 340). This use of the term is similar to that found in the literature on computer science and artificial intelligence where a heuristic is a prescription for action that is not guaranteed to work but is judged to be likely to do so.

The second way in which the term is used in the psychological literature is to connote reasoning strategies that often tend to lead the reasoner astray. Evans and Bradshaw (1986) have this connotation in mind when they refer to a heuristic as a theoretical construct that has been proposed to explain reasoning biases, or ways in which reasoning deviates systematically from the dictates of a normative statistical theory. Klahr (1976) likens heuristics of this type to "cognitive illusions," which are "compelling even when we know of their existence and can explain their source" (p. 245).

Kahneman and Tversky (1972b, 1973, 1982a; Tversky & Kahneman, 1973, 1974, 1983) have identified several heuristics that people appear to use when making judgments that call for probabilistic or statistical reasoning. The better known and more thoroughly studied of these are the availability, representativeness, and anchoring-and-adjustment heuristics. Whether these reasoning strategies should be considered heuristics in the first (beneficial) or second (detrimental) sense, or possibly both, is a question to which we will return. I note here that Kahneman (2000) takes exception to the widely held notion that people who do research on heuristics and biases are interested primarily in demonstrating human irrationality and argues that the greater interest is in "understanding the psychology of intuitive judgment and choice" (p. 682).

Availability

"Availability" is the name that Kahneman and Tversky (1973) gave to the heuristic that people use when they estimate the frequency of a class, or the likelihood of an event, on the basis of the ease with which they can bring examples or occurrences to mind. Kahneman and Tversky point out that availability can be a useful clue to relative frequency, because the relative frequency with which an event has been encountered is one of the determinants of its availability in memory; it is a fallible clue, however, because it is not the only such determinant. Availability may be affected also by such factors as vividness, concreteness, recency of activation, emotional quality, and other characteristics that enhance memorability (Billings & Schaalman, 1980; Shedler & Manis, 1986; S. E. Taylor, 1982), so its use can also lead to reasoning errors. In short, the availability heuristic can be expected to work well when the primary basis of availability is frequency of occurrence and to work less well when it derives from other factors; the problem is that its basis in specific instances usually is not clear. Much of the research relating to the use of the heuristic has focused on situations in which availability is likely to be influenced strongly by one or more factors other than relative frequency.

The use of the availability heuristic seems to account well for findings like the following ones reported by Kahneman and Tversky (1973). When people were asked to estimate the proportion of male or female names on a list of celebrities they had just read, their estimates were influenced by the relative fame of the people on the list. In particular, they tended to overestimate the representation of the gender which had the more famous names on the list, which, presumably, were the names they could more readily recall. In another example, people estimated the number of different 2-person committees that can be formed from 10 people to be greater than the number of different 8-person committees that can be formed from the same group. Although the number of possible committees is the same in both cases, the hypothesized explanation for the larger estimate in the two-person case is that two-person committees are easier to imagine than eight-person committees and are therefore judged to be more numerous.

Evidence regarding the importance of availability to likelihood estimates has been reported by S. J. Hoch (1984), who had people generate reasons why a future event might or might not occur. Some participants generated pro reasons first and then con, whereas others did the reverse. In both cases they generated more reasons of the type they were asked to produce first, and produced them more quickly. When asked to estimate the likelihood of the event, people's estimates tended to be consistent with the type of reason (pro or con) they had produced first. It appeared that the act of generating the initial list inhibited somewhat the generation of the second one and also tended to establish the participants' opinion on the matter. Increasing the length of the delay between the generation of the pro and con lists diminished the interference and increased the influence of the second list.

Availability seems to account also for the fact that when asked to judge the size of the set of English words that begin with a specified letter, relative to the size of the set of words that have that letter in the third-letter position, people tend to judge the set with the specified letter in the first-letter position to be the larger, even when there are many more words in the other set (Tversky & Kahneman, 1973). Presumably people can search memory more effectively when looking for words with a specified letter in first-letter position than when looking for words with that letter in third-letter position, and the greater availability of first-letter words supports the judgment that there are more of this class in the language.

The availability heuristic may be involved in a variety of other findings relating to systematic biases in estimation tasks. It could help account, for example, for why, when husbands and wives were asked to estimate the percentage of family activities (including unpleasant activities such as arguments) for which each spouse was responsible, the importance of each person's role was estimated higher by him or herself than by his or her spouse, or for why basketball players judged members of their own team to have been more responsible than members of the opposing team for critical plays in their games (M. Ross & Sicoly, 1979).

It appears also that people tend to overestimate the extent of their own contributions to tasks that they perform collaboratively with others (M. Ross & Sicoly, 1979). Honest misjudgments of this type are easily accounted for by the availability principle. One is bound to be more aware of the details of one's own work on a collaborative task, especially with respect to those aspects of the effort that are covert, than of the details of the work of one's collaborators. In recalling a project, one is likely to have available more information pertaining to one's own efforts than pertaining to the efforts of others, other things being equal.

Use of the availability heuristic has been considered one of the reasons why people often misestimate the frequencies of risky events; events that tend to be memorable, and hence available, tend to be perceived to occur with relatively high frequency (Slovic et al., 1981/1986). It also can help account for the common finding that people often tend to be overconfident of their own hypotheses. If, when generating hypotheses regarding the cause of an observed effect, people fail to think of some of the plausible alternatives to those they favor, the probabilities they assign to the latter are likely to be inappropriately high. People tend to produce less-than-complete sets of hypotheses and overestimate the likelihood of the hypotheses they produce (Fischhoff, Slovic, & Lichtenstein, 1978; Mehle, 1982; Mehle et al., 1981).

Representativeness

According to another hypothesis put forth by Kahneman and Tversky (1972b), people decide such questions as whether an object belongs to a specific category, or whether an event was generated by a specific process, on the basis of the degree to which the object or event is seen to be representative of objects in the category or events generated by the process. The subjective probability of an event or a sample, they argue, is "determined by the extent to which it: (i) is similar in essential characteristics to its parent population, and (ii) reflects the salient features of the process by which it is generated" (p. 430). An object or event can be representative of a category or process in a variety of ways (Tversky & Kahneman, 1983).

The idea has been applied not only to statistical reasoning but to categorization or classification more generally. Whereas natural categories were once defined by unique or distinguishing features, theorists today are more inclined to conceptualize them in terms of most-representative members (E. E. Smith & Medin, 1981; E. E. Smith, Osherson, Rips, & Keane, 1988). Representativeness, or typicality, is revealed by both direct ratings and indirect behavior indices, such as the time required to make a decision regarding category membership and the reliability with which such decisions are made. Members of a category that are more highly representative of the category are identified as members more rapidly and accurately than are members that are less representative—closer to the category's edges.

The representativeness heuristic has been invoked to account for some of the findings that have been described in connection with the "conjunction fallacy" (see chap. 9). The description of Linda, for example, might be said to be more representative of the class of feminist bank tellers than of that of bank tellers more generally. And, according to the hypothesis, use of the representativeness heuristic would lead one to select the more specific description over the more general one. The evidence seems to be that in many circumstances information that describes an individual as highly representative of a class has more influence on people's decisions regarding class membership than does base-rate information.

Kahneman and Tversky have also given a representativeness account of the fact that people are likely to judge the sequence of coin tosses HHTHTT to be more probable than the sequence HHHTTT, and much more probable than HHHHHH, despite the fact that all three sequences are equally probable. The explanation that invokes representativeness is that people expect even small samples selected at random from a large population to have the characteristics of the population from which they were drawn—this is one manifestation of the (fallacious) *law of small numbers*—and the first sequence is seen to be more representative of what one expects in a random sample of coin tosses than are the second and third.

Representativeness has been proposed as an explanation of why people often ignore or discount base rates in decision making, as in the widely discussed case of the blue and green cab problem. (see chap. 4.) This use of the concept has been criticized by Gigerenzer et al. (1989) on the grounds that it does little more than substitute one name of the phenomenon for another:

> The phenomenon is called base-rate neglect because people's judgments vary with $p(D|H)$ but not with $p(H)$, and this is explained by saying that people use a representativeness heuristic, which means that they use information of the type $p(D|H)$ and not $p(H)$. That is, neglect of base rates is explained by neglect of base rates. Probability theory has encompassed all. And the two concepts $p(H)$ and $p(D|H)$, from which the posterior probability is calculated, now serve as the vocabulary for both the phenomena and the explanations. (p. 224)

Gigerenzer and Murray (1987) claim that, inasmuch as representativeness is generally synonymous with likelihood, in attributing base-rate neglect to the use of a representativeness heuristic, one is, in effect, simply claiming that people use likelihoods instead of prior probabilities when updating probabilities. This may be true, they argue, but the question remains as to why they do so and use of the terminology of representativeness does not answer it.

Anchoring and Adjustment

It appears that people often make quantitative judgments by taking a tentative value as a point of departure and then making adjustments to it. Sometimes the starting point, or *anchor,* as it has been called, may be provided by someone else, or by some aspect of the context in which the judgment must be made. The primary finding of numerous experiments investigating the use of this heuristic is that when people are given an anchor, they typically adjust their judgments in the right direction but by an insufficient amount (Carlson, 1990; G. B. Chapman & E. J. Johnson, 1994; Slovic & Lichtenstein, 1971). It is as though they give more credence to the anchor than it deserves.

A demonstration of anchoring and (insufficient) adjustment is reported in Tversky and Kahneman (1974). One group of participants was asked to state the probability that the population of Turkey was greater than 5 million and another group was asked to state the probability that the population of Turkey was less than 65 million. Later, when each group was asked to estimate the population of Turkey, the median estimate for the first group was 17 million and that for the second group 35 million. Given that the only known relevant difference between the groups was the fact that they were exposed to different numbers when asked the first question, it seems reasonable to conclude that that difference was instrumental in determining the difference in their subsequent estimates. In anchoring-and-adjustment terms, we might say that the initial numbers provided by the experimenter served as the anchors for the participants' subsequent estimates and the differences between the initial numbers and the estimates themselves constituted the adjustments that were made.

Without contesting the appropriateness of this terminology, we can note that using the initial number provided by the experimenter as an anchor for the subsequent population estimate could be a sensible thing for one to do. Suppose one had no idea what the population of Turkey is. It would not be unreasonable to assume that, in order to make the first problem difficult, as they might be expected to do, experimenters would choose numbers fairly close to the actual population of Turkey. On this line of reasoning, people who came to the experiment with no idea of Turkey's population might do worse than take the number provided by the experimenter as a default point of departure when asked to produce an estimate.

A similar point could be made with respect to a finding by Lichtenstein et al. (1978). In this case people were asked to estimate the frequency of fatalities associated with each of 40 causes of death in the United States. Some, who were told by the experimenters that there are about 50,000 highway fatalities each year, gave much higher estimates than did others, who were told that there are about 1,000 deaths due to electrocution. Again, individuals who came to the experiment without any clear idea about the frequencies of deaths from accidental causes might have taken the numbers provided by the experimenters to be "typical" in some sense.

The same rationalization of people's behavior is not so easy to apply to another experiment of Tversky and Kahneman's (1974), however, in which students were asked to estimate the number of African countries in the United Nations. In this case, after making an initial estimate, the students watched a roulette wheel being spun, and were asked, first, if the number they had estimated was above or below that which the wheel generated and, second, to give a new best estimate. The final estimate was affected by the number produced by the wheel, even though students actually saw the wheel spin and presumably knew that it was a random process.

In still another experiment, Tversky and Kahneman (1974) asked two groups to estimate the product of a series of numbers quickly. One group saw the sequence 1 x 2 x 3 x 4 x 5 x 6 x 7 x 8, and the other 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1. Both groups gave estimates that were erroneously low, but the estimates of the first group were substantially smaller than those of the second. An account of these results in terms of anchoring and adjustment would assume that both groups estimated a partial product of the first few terms and, using that as an anchor, made adjustments to accommodate the remaining terms. The underestimation by both groups could be attributed to insufficient adjustment; the fact that the first group underestimated by a larger amount than the second follows from the first group starting with a smaller anchor. Anchoring and adjustment in this context seems like a perfectly reasonable way to approach the problem; the difficulty appears to be not with the process but with a lack of accuracy at the various steps.

Anchoring and (insufficient) adjustment has also been used to account for the finding, already noted, that when people know the probabilities of two independent events, they often overestimate the probability of the conjunction of those events and underestimate the probability of the disjunction (Tversky & Kahneman, 1974). In the case of the conjunction, the argument is that the anchor is the probability of one of the events and the adjustment (downward) for the conjunction is too small. In the case of the disjunction, the anchor again is the probability of one of the events and the adjustment (upward) for the disjunction is again too small.

Another possible example of anchoring and adjustment involving estimation comes from a study by Lopes and Ekberg (1980). When evaluating gambles (relative to "sure thing" alternatives) people made the judgments somewhat faster when the amount to be won was presented before the probability of winning, as opposed to the reverse. This was interpreted as suggesting that the gambles were evaluated by using the value of the amount to be won as an "anchor" and then adjusting this down to reflect the probability of winning.

A final example of what might be viewed as a case of anchoring and adjustment is a considerably more troubling one. It involves a study by Hamil, T. D. Wilson, and Nisbett (1980), in which people were shown one of two videotaped interviews of an actor posing as a prison guard. In one of the interviews the "guard" was compassionate and appeared to be interested in prisoners' rehabilitation; in the other he was verbally abusive of prisoners and showed little interest in their well-being. Subsequent testing of the participants showed that their beliefs about prison guards in general had been strongly influenced by the videotape viewings, even when they had been told that the guard they had seen on tape was exceptionally humane (or inhumane) and not representative of guards in general; when asked to estimate attitudes typical of prison personnel on a variety of issues, participants estimates appeared to have been influenced as much by the expressed opinions of guards who had been characterized as extreme as by those who had been characterized as typical. This result suggests that opinions and attitudes may sometimes be anchored by information we receive even when there is reason to believe that information to be distorted or biased in some way.

Other studies that have found evidence of adjustment from an anchor include those of Cervone and Peake (1986), H. L. Davis, S. J. Hoch, and Ragsdale (1986), E. J. Johnson and Schkade (1989), and Schkade and E. J. Johnson (1989). Dawes (1988) points out that the most common anchor we have—and we always have it—is the status quo. Perhaps this is why we find it easier, usually, to imagine variations on existing themes than radically new themes.

Anchoring and adjustment may be seen in contexts other than those involving statistical reasoning in the usual sense. People who deal in antiques, objects of art, memorabilia, and other goods the prices of which are not established by the dynamics of mass production and competition between producers of nearly identical products in the marketplace understand the principle very well. When the dealer is the seller, the point of departure for dickering is the initial asking price. The unsophisticated buyer is likely to consider a purchase at a large (say 25%) discount from the asking price to be a bargain, but this involves the assumption that the asking price was reasonable. If the object involved is unique or relatively so, it may be very difficult to determine what is "reasonable" in an objective way, and the dealer who intuitively understands the anchoring and adjustment principle does well to begin with a high price. (When the dealer is the buyer, dickering starts with the initial offer, and the same principle holds in reverse; in this case, the dealer does well to begin with a low price and make the seller feel good by raising it by some significant fraction.)

I have argued that anchoring and adjustment may play a role determining our assumptions about what other people know on particular subjects (Nickerson, 1999, 2001). In this case the anchor is what one knows, or thinks one knows, oneself. This anchor serves us well for the most part, but often the adjustment that is made to take account of individual differences in knowledge is too small and the result is that we overestimate the probability that a person has a particular bit of knowledge that we ourselves have and effective communication is impeded.

Whatever conclusion one draws regarding the adequacy of the anchor-and-adjustment explanation for many of the results that have been obtained in experiments on reasoning under uncertainty, the evidence of the reality of the phenomenon is compelling. An important practical implication of the demonstration of its existence is the clear revelation that the answers one gets to questions can depend very much on precisely how the questions are posed. This is perhaps something that everyone knows in a qualitative way from personal experience, but it is useful to have such striking evidence of the effect and indications of the nontrivial magnitudes it can reach in specific cases.

Usually the phenomenon of anchoring and adjustment is treated as a type of bias in reasoning and an example of one of the many ways in which thinking is less than completely rational. G. B. Chapman and Bornstein (1996) make the important point, however, that in many cases, the anchor provides information that is relevant to the reasoner's task. And even when it does not, if participants in experiments assume that it does, they may be applying a principle that generally holds in normal discourse (Grice, 1975).

How Effective Are These Heuristics?

> All heuristics make us smart more often than not, and all heuristics—by mathematical necessity—induce weighting biases. (Kahneman, 2000, p. 683)

There can be little doubt that people use the types of heuristics identified by Kahneman and Tversky when making judgments and decisions about statistical or probabilistic variables. As Tversky and Kahneman (1974) point out, these principles simplify the task of reasoning about uncertain situations and, in general, are quite useful. They frequently yield correct conclusions or at least conclusions that are correct enough for the practical demands of the moment. However, as Tversky and Kahneman also point out, they "sometimes lead to severe and systematic errors" (p. 1124).

Piatelli-Palmarini (1994) puts the case more strongly:

> We have come to see that our minds spontaneously follow a sort of quick and easy shortcut, and that this shortcut does not lead us to the same place to which the highway of rationality would bring us. Few of us suffer from any illusion that the summary paths taken by our intuitions and approximations would lead us to *exactly* the same point to which reason and exact calculation might have brought us. But we do delude ourselves into thinking that we are thereby brought to a neighboring area, one that is *close enough.* (p. 143)

It seems clear that the use of the kinds of heuristics being considered sometimes yields judgments and decisions the rationality of which appears dubious at best. What should we conclude about them in general? Is their use stark evidence of human irrationality? Should we try, through the educational system, to teach students not to use them?

We do not know how pervasively these heuristics are used outside the psychological laboratory. Several investigators have questioned whether that use is really very extensive, and not all are convinced that the types of problems and problem contexts that have been used to study them in the laboratory are representative of those that people typically encounter in daily life (L. J. Cohen, 1981; Dennett, 1981; Lopes, 1982; Macdonald, 1986).

We also do not know if, when they are used outside the laboratory, they typically work well or poorly. Many of the experimental situations in which they have been studied have been carefully designed to demonstrate how they can lead one astray. But how representative are these contrived situations with those people typically encounter in everyday situations? One of the criticisms of laboratory demonstrations of cognitive illusions is that experimenters sometimes get the appearance of irrational behavior only by violating widely accepted principles of normal discourse in their structuring of problem scenarios. If heuristics of the type considered here are used extensively, and their use typically results in faulty judgments and decisions, why have they survived? Would we not expect grossly ineffective approaches to judgment and decision making to become casualties of natural selection over time?

One can make a case for the practical utility of these heuristics, at least under certain conditions. As already noted, availability might be expected to be correlated with relative frequency and therefore often to serve as an effective clue to relative frequency in statistical reasoning. The correlation is not perfect, of course, and availability is known to be affected by factors other than relative frequency, so its use as an indicant of relative frequency can lead to reasoning errors; but it is at least conceivable that availability is a sufficiently reliable indicator of relative frequency to ensure the availability heuristic's usefulness in many contexts.

J. R. Anderson (1990) has pointed out that the utility of any heuristic approach to a cognitive problem should be evaluated in terms of the expected value of the consequences of applying it. To borrow the example he uses to make the point, the tendency that some people have not to believe an argument if the arguer appears not to believe it is normatively irrational, inasmuch as the validity of an argument does not depend on the beliefs of the one who advances it, but it is possible that people who use this heuristic are less likely to accept an invalid argument than those who do not.

Gigerenzer (1991; Gigerenzer & Murray, 1987) has criticized the use of the Kahneman and Tversky heuristics to account for errors in probabilistic reasoning on several grounds. First, he argues that many of the "errors" that they are invoked to explain are not really errors, or are so only from a particular narrow and challengeable interpretation of probability. Second, he sees the heuristically based explanations, in several cases at least, as little more than redescriptions of the phenomena they are intended to explain. Third, he dismisses the heuristics considered previously as useful explanatory constructs for much the same reason that Popper (1959) dismissed the theories of Marx, Freud, and Adler, namely, their excessive versatility: These heuristics, in his words, "are largely undefined concepts and can *post hoc* be used to explain almost everything" (p. 102).

My opinion regarding heuristic approaches to intellectually demanding tasks is that they are valuable to the extent that they (a) work acceptably well in most of the important situations in which they are applied, and (b) make significantly lighter demands on one's cognitive resources than alternative approaches that would give more precise or more consistently correct results. What one gains by using heuristics is computational simplicity, the ability to address complex problems with only modest effort; what one gives up is precision and guaranteed statistically optimal results. Is this a reasonable trade?

I doubt if it is possible to answer this question in any very conclusive way. We can, however, reflect on the question and on what some partial answers might be. It is conceivable, for example, that the trade is worth it in the sense that the ratio of the benefit derived from the use of heuristics to the cost of their use is greater, when integrated over all situations in which they are used, than would be the ratio of the benefit derived from the use of more exact approaches to the cost of their use. It also could be that heuristics work well enough in situations that really matter to offset their ineffectiveness in less important situations. Both of these possibilities could be seen as adaptive.

The question arises, if a heuristic procedure is effective in some situations and not in others, why is it not applied only in situations of the former type. There are at least two plausible answers to this question. The first challenges the tacit assumption on which the question rests. There may well be examples of heuristics that are used to advantage on some occasions by some people and avoided in favor of more exact methods on other occasions by the same people. We do not know that this is not the case. A second plausible answer addresses those cases in which a heuristic procedure is used more or less invariably, assuming there are such. Invoking again notions of benefits and costs, one might assume that what is to be gained by deciding on a case-by-case basis whether use of the heuristic is likely to yield acceptable results or a more precise approach is called for is not worth the cost of being constantly faced with the need to make this decision.

This is all conjectural, but some conjecture is probably required to see the use of heuristics from a sufficiently broad perspective. The evidence is compelling that situations can be designed in which people apply heuristics and, as a consequence of doing so, form judgments, draw conclusions, or make decisions that are suboptimal in specifiable ways. When viewed in isolation, such applications of heuristics are likely to be taken as instances of irrational behavior, but when viewed from a broader perspective these applications might be seen as exceptions to the general usefulness of the procedures involved.

None of this is to deny the possibility of making people more aware of the benefits and limitations of heuristic thinking and more discriminating in the use of specific heuristic rules. Presumably not all heuristics are equally effective, relative to any measure of effectiveness that might be applied, and specific heuristics are applied to better advantage in some situations than in others. It is hard to believe that a more extensive and explicit knowledge of the advantages and limitations of heuristic thinking and of specific heuristic rules could be detrimental to one's rationality and it is easy to see how it might increase the effectiveness with which one could meet intellectual challenges in general.

PROBABILISTIC MENTAL MODELS AND ANOTHER VIEW ON HEURISTICS

Gigerenzer (1991, 1993) and his colleagues (Gigerenzer et al., 1991) have proposed a theoretical account of how people reason about uncertain situations that combines the concept of mental models with the assumption that people naturally distinguish between frequencies and other meanings of probability. According to the theory, people deal with uncertain situations by constructing probabilistic mental models (PMMs) of them: "A PMM is a generalization of the particular problem, and consists of a reference class of objects and a network of probability cues" (Gigerenzer, 1991b, p. 104). Gigerenzer and colleagues have also engaged in a program of research aimed at identifying simple ("fast and frugal") heuristics that take advantage of the structure of information in the environment and lead to fast and relatively accurate decisions with a minimum expenditure of cognitive capital (Gigerenzer & Todd, 1999a).

Probabilistic Mental Models

To illustrate the operation of a PMM, Gigerenzer uses the general-knowledge question: "Which city has more inhabitants? (a) Heidelberg, (b) Bonn." To derive an answer to this question—assuming one did not know the answer but had some knowledge of German cities—one would generalize Heidelberg and Bonn to a reference class, such as all German cities, and number of inhabitants to a network of probability cues that might include such specifics as whether one city, but not the other, has a professional soccer team or is a state capital. Probability cues differ in degree of validity and one's perception of a cue's validity is based on learned frequencies of co-occurrence. A judgment regarding a question like the one relating to Heidelberg and Bonn would be made on the basis of having found one or more applicable probability cues (not all cues relevant to city size would be applicable to a given pair of cities) and one's confidence in it would depend on the subjective validity of the cue(s) activated. Because "PMM theory postulates cognitive mechanisms that work well given limited knowledge, limited attention, and limited computational capacities" (p. 301), it is seen by Gigerenzer (1993) to be a model of "bounded rationality" of the sort described by Simon (1955). The process mentioned produces good, but not necessarily optimal, performance; it is therefore a satisficing process in Simon's sense.

Gigerenzer invokes the PMM hypothesis to account for the fact that people give conflicting responses when asked to estimate the probability that an answer to a particular question is correct and when asked to estimate the percentage of the last, say 50, questions they have answered correctly. He suggests that the PMMs that people construct in the two situations differ in critical ways. In particular, in the former case the target variable, reference class and probability cues are determined by the subject of the question, whereas in the latter the target variable is "number of correct answers," the reference class may be similar testing situations, and the probability cues may be performance results in those situations.

A particularly interesting prediction that Gigerenzer (1993) makes from PMM theory is that limited knowledge may sometimes provide as good a basis for judgment as more extensive knowledge. He argues the plausibility of this prediction with the following illustration. Consider the task of judging for each of 100 pairings of the 75 largest cities in a country which is the larger. Suppose the same people—let us say a group of German students—are asked to perform this task once with pairs of German cities, with which they are likely to be highly familiar, and once with pairs of U.S. cities, about which they are likely to know much less. In making the judgment with respect to any given pair of German cities, several probability cues are likely to be accessible, some of which may point to one of the cities as the larger and others of which may point to the other. When dealing with any given pair of U.S. cities, one is likely to have fewer cues available; in some cases familiarity—whether or not one has heard of the city—may be the only one. If the available cue is highly indicative of relative size, however, it may be as useful as would a large set of conflicting cues in helping one make the right choice. Gigerenzer is able to show that, given some plausible assumptions about cue accessibility and validity, it would not be surprising to find the German students doing as well at this task with U.S. cities as with German ones, and he reports some data collected with colleagues showing that a group of German participants did essentially equally well with German and U.S. cities on this task.

Fast and Frugal Heuristics

The idea that sometimes lack of knowledge can be advantageous by actually increasing the probability of selecting the correct item among two or more alternatives considered has been pursued experimentally by Gigerenzer and colleagues. D. G. Goldstein and Gigerenzer (1999) describe the results of experiments showing that basing choices on simple recognition can be an effective strategy in dealing with situations in which one knows little or nothing about the alternatives, but recognizes (only) one among them. This heuristic is domain specific, Goldstein and Gigerenzer caution, working only in domains in which recognition is correlated with the property with respect to which the choice is to be made. It works in the case of selecting the foreign city among two or more alternatives that has the largest population, presumably because the greater the population of a city, the greater the probability that one has heard of it.

In a particularly thought-provoking study, Borges, D. G. Goldstein, Ortmann, and Gigerenzer (1999) showed that investments in stocks selected solely on the basis of recognition of company names outperformed stock portfolios selected on the basis of conventional criteria, in some cases by experts. The investigators conclusion that "In general it seems that the greater the degree of ignorance, the better it is for picking stocks" undoubtedly needs some qualification; else we would all be millionaires, but their account of how lack of knowledge can be beneficial in specific situations is instructive.

Gigerenzer and D. G. Goldstein (1999) consider several heuristics for deciding when to stop searching for cues on which a selection might be made in a choice situation and then making the selection. Suppose several cues are available regarding the property of interest; cues to the size of a European city, for example, might include whether it has a soccer team, whether it is a national capital, whether it is home to a university, ... If the cues can be ordered in terms of their correlation with the property of interest, one often may do quite well by relying solely on the single best cue among those one knows about. It works best if, when the cues are ordered in terms of their importance, each cue is more important than any combination of the less important cues (Martignon & Hoffrage, 1999). Gigerenzer and Goldstein call this heuristic *Take the Best,* and they present data showing that it can be about as effective as standard statistical analyses, including multiple regression (which require much greater computational capability) in specific instances.

Other studies of the effectiveness of fast and frugal heuristics and single-reason decision making—and especially of the Take the Best heuristic— are reported by Czerlinski, Gigerenzer, and D. G. Goldstein (1999), Martignon and Hoffrage (1999), Martignon and Laskey (1999), and other contributors to Gigerenzer and Todd (1999b). Heuristics are described for estimating quantities (Hertwig, Hoffrage, & Martignon, 1999), categorizing objects (Berretty, Todd, & Martignon, 1999) or actions (Blythe, Todd, & Miller, 1999), searching for a mate (Todd & G. F. Miller, 1999), and investing by parents in their offspring (J. N. Davis & Todd, 1999). The general idea promoted by these investigators is that people generally reason by availing themselves selectively of the contents of a toolbox of ecologically rational simple heuristics that are effective despite making limited demands on memory and computational resources. The heuristics in the toolbox have been adapted over time and they are retained because they work—in some cases as well as, or even better than, formal procedures that require significant computational resources. Different tools are suited to different situations and it is important that there be a match between a tool that is selected and the structure of the situation in which it is to be applied. A major challenge for this line of research is that of identifying the conditions under which specific tools should be used (Luce, 2000) and of safeguarding against applications that can have the opposite of the desired effect (Margolis, 2000).

SOME POTENTIAL PITFALLS

Several variables that can influence the effectiveness with which people make judgments about probabilistic events have been identified by researchers. Because the effects of these variables usually, though not always, are detrimental, at least as observed in psychological laboratories, I am discussing them under the loose rubric of potential pitfalls. "Potential" is a considered qualifier in this heading, however, because in at least some of these cases, what can lead to the pit in one context can be an effective strategy in another.

Framing Effects

According to Kahneman and Tversky (1984), "all analyses of rational choice incorporate two principles: *dominance* and *invariance*. Dominance demands that if Prospect A is at least as good as Prospect B in every respect and better than B in at least one respect, then A should be preferred to B. Invariance requires that the preference order between prospects should not depend on the manner in which they are described" (p. 343). Kahneman and Tversky review several evidences that the requirement of invariance is often violated by human choice behavior. In particular, the preferences that people have in choice situations may be determined by the way in which the options are presented or "framed," even when the framing has no effect on the choice outcomes. The following alternative framings of a problem, from Kahneman and Tversky (1984), illustrate the point:

> Problem: Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> Frame 1: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. Which of the two programs would you favor?
>
> Frame 2: If Program C is adopted, 400 people will die. If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die.

Kahneman and Tversky (1984) found that 72% of 152 people who were given the first version of this problem selected Program A, whereas 78% of 155 people who were given the second version selected Program D. The same type of result was obtained with sophisticated participants as with naive ones, and even when the same people responded to both representations within a period of a few minutes. Kahneman and Tversky's explanation invokes the notion that people tend to be risk aversive when thinking in terms of gains and risk seeking when thinking in terms of losses. The first representation of the epidemic problem formulates the outcomes of the alternative programs in terms of two possible gains relative to a reference point in which 600 people are expected to die. The second representation provides options expressed as losses relative to a reference point in which no one dies.

This is a compelling demonstration that preferences among decision outcomes are determined not only by what the outcomes are, but by the way in which they are represented. The result has been replicated in other studies not only when the differently framed choices were given to different groups of people, but also when the same people were given both types of frames (Dawes, 1988). However, before accepting without reservation the conclusion that people's *preference for a given outcome* changes as a function of how that outcome is described, we would like to feel sure that they believe that an outcome described in different ways is indeed the same outcome. In the aforementioned case, we would want to be convinced that people believed the outcomes of A and C (and of B and D) would be the same. Without this assurance, we cannot rule out the possibility that the their choices were determined by what they *thought* the outcomes would be, as opposed to being influenced by how identical outcomes were expressed.

The intent of the experimenters was to frame precisely the same problem—present precisely the same information—in different ways. The assumption is that the assertion "200 people (out of 600) will be saved" is equivalent to the assertion that "400 people (out of 600) will die." As Macdonald (1986) points out, however, such assertions, as they are used in everyday language, may not be seen as conveying exactly the same information. In many contexts, a guarantee to save 200 people could be taken as a guarantee to save *at least* 200. If more than 200 were saved, the guarantor would have been seen as keeping his promise, whereas, if less than 200 were saved he would not. Similarly the assertion that 400 people will die could be taken as a threat to 400 lives at least. If the language used in the problem is interpreted this way, then A and C are not equivalent and do not describe equally desirable choices: "Subjects may see themselves as comparing an uncertain result with either a clearly positive one (A) or a clearly negative one (C)" (p. 24). If linguistic ambiguities of this sort play some role in determining people's choices, we still might want to describe the effects as framing effects, but we would have to acknowledge the possibility that framing effects result, at least in part, from the fact that different frames may lend themselves to linguistic interpretations that are not equivalent in meaning. Others have also argued this possibility (Berkeley & Humphreys, 1982; Kühberger, 1995).

Even if the role of linguistic ambiguities were not an issue in interpreting the framing effect described, it is not clear that the people's behavior in this situation should be considered irrational. J. R. Anderson (1990) points out that, although Kahneman and Tversky's (1984) theory of problem framing accounts for the choices their subjects made, there is little reason to consider those choices irrational; inconsistent behavior should not be considered evidence of irrationality in a situation in which there is no good reason to be consistent, and no rational model would prescribe a consistent preference between choices with equivalent outcomes. Anderson notes also that in some other choice situations in which framing effects have been found, the stakes riding on the choice have been small. Inconsistent behavior in such instances, he argues, may constitute evidence against rationality in the normative sense, but make a weak case against it in the adaptive sense.

Different ways of presenting essentially the same information can have different effects on those to whom the information is given. The relative attractiveness of different medical procedures, for example, differs for both physicians and patients depending on whether the probable outcomes are described in terms of mortality or survival (McNeil, Pauker, Sox, & Tversky, 1982) and on precisely how the degree of uncertainty regarding the outcome is expressed (Teigen & Brun, 1999). In betting situations, choices vary depending on whether a particular gamble is expressed in terms of probability of winning or probability of losing (I. P. Levin, D. P. Chapman, & R. D. Johnson, 1988; I. P. Levin et al., 1986; I. P. Levin, R. D. Johnson, Russo, & Deldin, 1985). Lobbyists for the credit card industry have shown a sensitivity to framing effects in arguing that price differences between cash and credit purchases be expressed as cash discounts rather than credit card surcharges (Thaler, 1980).

How such results should be interpreted depends in part on the confidence one can have that the two forms of a given message have been equally understood by those to whom they were given. The point is illustrated by an experiment in which people were either (a) informed that over a period of 50 years of driving (about 40,000 trips) the probability of being killed is about 1 in 100 and that of experiencing at least one disabling injury is about 1 in 3 or (b) given the equivalent information by being told that there is one fatal accident for every 3.5 million person trips and a disabling injury for every 100,000 person trips. Those in the former group responded more favorably to the use of seat belts than did those in the latter (Slovic et al., 1978). Perhaps most people would find the first message somewhat easier to relate to their personal chances of having a serious accident, simply because it is phrased explicitly in those terms. Similar questions of comprehension can be raised about many of the studies in which presumably the same information was conveyed in two different ways.

Toda (1963) has argued that subjective probability is, in effect, defined by the technique that is used to measure it. This argument gets some support in the finding that the way people are asked to give probability estimates affects the estimates they produce (Damas, Goodman, & C. R. Peterson, 1972; Herman, Ornstein, & Bahrick, 1964; Pitz, 1974; Seaver et al., 1978).

A powerful effect of question wording was obtained by Fischhoff and MacGregor (1983) (reported also in Fischhoff, Slovic, & Lichtenstein, 1982, and in National Research Council, 1989, Appendix C). When asked to estimate the lethality rate from influenza people estimated about 393 deaths per 100,000 influenza cases. When told that about 80 million people have influenza in a normal year and asked to estimate the number of those cases that would end in death, their average estimate was 4,800, which is equivalent to only about 6 deaths per 100,000. Here, too, one might assume that in telling people how many cases of influenza there are in an average year, one is providing information that they might consider relevant to the rate estimate; applying the rate of 393 per 100,000 to a total case number of 80 million would produce an estimate of more than 300,000 deaths, which might seem implausibly high.

There is abundant evidence that public opinions, as reflected by the results of polls, can vary considerably depending on the way in which questions are framed (Moore, 1992; Payne, 1952; Wheeler, 1976). Even the degree of satisfaction with one's life that one expresses can depend on the range of options provided (Parducci, 1974). Especially problematic is the use of words that can have a variety of connotations. It is hard to know, for example, what to make of the results without knowing how people interpret such terms as *possible* and *doubt*. Some people may apply *possible* only to events that they consider to be somewhat likely, whereas others may apply it to those they consider even remotely conceivable; similarly, *doubt* can connote anything from fairly strong disbelief—"Tom believes that Elvis is alive and well, but I really doubt it"—to recognition of the remote possibility of being wrong—"I am convinced this is the right thing to do, although I confess to a lingering doubt."

The Endowment Effect

It is fairly easy to get people to make what appear to be inconsistent choices under uncertainty. An example from Thaler (1980; 1983/1986) will make the point. Thaler (1983/1986) describes the following three situations:

> Risk Situation 1: While attending the movies last week you inadvertently exposed yourself to a rare, fatal disease. If you contract the disease, you will die a quick and painless death in one week. The chance that you will contract the disease is exactly .001—that is, one chance in 1000. Once you get the disease there is no cure, but you can take an inoculation now which will prevent you from getting the disease. Unfortunately, there is only a limited supply of inoculation, and it will be sold to

> the highest bidders. What is the most you would be willing to pay for this inoculation? (If you wish, you may borrow the money to pay at a low rate of interest.)
>
> Risk Situation 2: This is basically the same as situation 1 with the following modifications. The chance you will get the disease is now .004—that is, four in 1000. The inoculation is only 25 percent effective—that is, it would reduce the risk to .003. What is the most you would be willing to pay for this inoculation? (Again, you may borrow the money to pay.)
>
> Risk Situation 3: Some professors at a medical school are doing research on the disease described above. They are recruiting volunteers who would be required to expose themselves to a .001 (one chance in 1000) risk of getting the disease. No inoculations would be available, so this would entail a .001 chance of death. The 20 volunteers from this audience who demand the least money will be taken. What is the least amount of money you would require to participate in this experiment? (p. 163).

Thaler reports that people to whom he has presented these situations typically give median responses of about $800 in Situation 1, $250 in Situation 2, and $100,000 in Situation 3. There is a sense in which what one is buying in the first two cases is the same, namely a reduction of .001 in the likelihood of contracting a disease, which is equivalent to what one is selling in the third situation. People's responses here certainly appear to be inconsistent. In particular, comparing the responses to Situation 3 with those to Situations 1 and 2, we see a very large difference. In Thaler's (1983/1986) words, "this implies that a typical individual would refuse to pay $5,000 to eliminate a risk, *and* would refuse to take $5,000 to accept the same risk. How can $5,000 be both better and worse than bearing some risk?" (p. 164).

It appears from these results that people are not willing to pay as much to decrease an existing risk as they require to be compensated for accepting a new risk of comparable magnitude. Thaler refers to this phenomenon as the "endowment effect," the idea being that people demand more money to give something up than they are willing to pay to acquire it. The effect has been observed in other contexts than that just described (Kahneman, Knetsch, & Thaler, 1990).

To illustrate the endowment effect in a situation that does not involve risk, Thaler poses the following decision situation: "Suppose you won a ticket to a sold-out concert that you would love to attend, and the ticket is priced at $15. Before the concert, you are offered $50 for the ticket. Do you sell? Alternatively, suppose you won $50 in a lottery. Then a few weeks later you are offered a chance to buy a ticket to the same concert for $45. Do you buy?" According to Thaler many people say they would neither sell for $50 in the first case nor buy for $45 in the second. A practical implication that is suggested by the endowment effect is that people may be willing to work harder to hold on to something they have than to acquire it in the first place. Or to put it in other words, having acquired something, people may be very reluctant to give it up, even if they did not greatly desire it originally.

We have already noted that people often express a preference for a gamble with a relatively small potential payoff and a relatively large probability of winning over one with the same expected payoff but composed of a larger potential payoff and smaller probability of winning. That is they opt for the larger probability of winning. If given the chance to sell either gamble, however, the same people are likely to demand a higher price for the second one than for the first (Grether & Plott, 1979). One account of this apparent inconsistency is that when the focus is on winning versus losing, people anchor on the probability of success, whereas when the focus is on selling, they anchor on monetary amounts, and use of the anchor-and-adjustment heuristic in the two cases yields different results because it starts from different points (Dawes, 1988).

An endowment effect of sorts is easy to understand in intuitive terms in some contexts. I have a small collection of antique wood-working tools. There are a few tools in that collection that I would be unwilling to sell even if offered considerably more than I believe to be their fair market value, but I did not pay more than what I considered their fair market value to be when I acquired them, and if I did not already have them, I would be unlikely to do so now. The value of these tools to me is determined, in part, by the fact that while in my possession they have become to me more than objects. They represent memories of successful treasure hunts with my wife and of many pleasant hours spent learning about their functions, working with them, and admiring the craftsmanship that produced them. Some have special significance because of unusual places or circumstances in which they were found.

I suspect that most of us, if asked, would be able to identify things we possess that have value to us by virtue of the fact that we possess them. We would be unlikely to be willing to sell them for some specified amount of money that is greater than what we would be willing to pay for them if we did not possess them already. In fairness to those investigators who have seen the endowment effect as evidence of an irrational inconsistency in choice behavior, I think that instances of personal possessions that have come to have special significance to their owners are not the types of entities for purchase and sale they have had in mind. My point is that one can easily think of circumstances under which it seems reasonable and not surprising that the desire to hold on to something one has is greater than the desire to acquire it in the first place; I am not prepared to argue that this asymmetry is reasonable in all cases.

Admissible Scoring Rules

Considerable attention has been given by researchers to the fact that some of the techniques that have been used to get probability estimates from people have the property that it is not in one's best interest to report one's real subjective probabilities. The problem is illustrated by the following situation. Imagine a test composed of two-alternative forced-choice questions. Suppose that, instead of selecting one of the alternatives in each case, the task is to state, for each alternative, one's subjective probability that it is correct. (Assume that one and only one of the alternatives is correct in each case, so one's subjective probabilities for each pair of alternatives must sum to 1.0.) Suppose that when the test is scored, the score received for each item is the amount placed on the correct alternative, or some linear function thereof. Thus, for example, if a probability of .8 had been assigned to the correct alternative on Item 1 and one of .4 to the correct alternative on Item 2, .8 and .4 would be credited respectively for these items, or, in any case, twice as much for the first item as for the second. It is easy to show that, given this scoring rule, one should *not* assign numbers to the alternatives in such a way as to reflect one's true beliefs about their probabilities of being correct; rather one should assign 1.0 to any alternative whose probability one considers to be greater than .5, and 0 to any whose probability one considers to be less than .5. (For those one considers to be exactly .5, one could assign .5 or either 1.0 or 0 on the basis of a toss of a coin.)

To see why the "all-or-none" strategy is better than that of weighting the alternatives according to one's real subjective probabilities, consider all those cases in which one believes the probability of a given alternative is .7. If .7 is assigned to those alternatives, then one expects to have assigned this value to the correct alternative 70% of the time and to the wrong alternative the remaining 30%. Thus, one's expected average score for this subset of items is (.7)(.7) + (.3)(.3), or .58. On the other hand, if 1.0 is assigned to all those items that one believes to be correct with probability .7 and 0 to those one believes to be correct with probability .3, one's expected average score for these items is (.7)(1.0) + (.3)(0) or .70. One does better in the latter case. (The situation corresponds to the one discussed relative to probability matching in chap. 8.)

Several scoring rules have been invented that have the property, sometimes called the "matching property," that one maximizes one's expected score when one reports one's true subjective probabilities. These rules, which have been referred to as "admissible probability measures" and as "proper scoring rules," have been the subject of both theoretical and empirical investigation (de Finetti, 1962; Good, 1952; Roby, 1965; Shuford, Albert, & Massengill, 1966; Toda, 1963). I will describe briefly one such rule—the spherical-gain rule—to illustrate the concept.

According to the spherical-gain rule, described by Roby (1965), one's score on any question (in a multiple-choice test item for which one and only one answer is correct) is the number assigned to the correct alternative divided by the square root of the sum of squares of the numbers assigned to all the alternatives; that is,

$$s_j = x_{j,c}\left(\sum_k x_{j,k}{}^{2}\right)^{-\frac{1}{2}},$$

where s_j represents the score received for the jth question on the test, $x_{j,k}$ the number that the student assigns to the kth alternative for the jth question, and $x_{j,c}$ the number assigned to the correct alternative for that question. (The rule gets its name from the fact that the assignments of numbers to n alternatives can be represented as vectors in an n-dimensional space and interesting interpretations can be given to various properties of the vectors, such as length and orientation.)

Imagine, for illustrative purposes, the situation represented in Table 10.1. It should be obvious that $0 \le s_j \le 1$. The score will be 0 if 0 has been assigned to the correct alternative (as in A above); it will be 1 if 0 is assigned to every alternative except the correct one (as in B). A "pure guess" in which the same nonzero number is assigned to all alternatives (as in C) will result in a small, but not 0, score. (This feature reflects the idea that knowing that one does not know is

TABLE 10.1

Question: Which of the Following U.S. Presidents Served Two Nonconsecutive Terms?

1. James Madison
2. Grover Cleveland
3. William Harding
4. John Adams
5. James Buchanan

(The correct answer is #2.)

Hypothetical answers:

A.	1	0	B.	1	0	C.	1	1	D.	1	0
	2	0		2	10		2	1		2	7
	3	0		3	0		3	1		3	0
	4	5		4	0		4	1		4	6
	5	0		5	0		5	1		5	2
Scores	0.00			1.00			0.45			0.74	

preferable to believing that one knows when one really does not.) Answer D represents a case in which one can rule out two alternatives and is able to express differential confidence about the remaining three. The score in this case rewards the test taker for the partial knowledge displayed.

In general, the larger the number placed on the correct alternative relative to the numbers assigned to the other alternatives, the larger the resulting score. More important, one maximizes one's expected score if (and only if) one assigns numbers to the alternatives in accordance with one's true beliefs regarding their relative chances of being correct. Consider again the two-alternative case mentioned earlier in which one believes that the probability that a particular alternative is correct is .7. As noted previously, given a linear scoring rule, one's best strategy is to put all one's bet on the most likely alternative, in this case the one with subjective probability .7. If one bet in accordance with one's beliefs one's expectation would be .58, whereas if one put all one's bet on the most likely alternative, it would be .70.

To simplify the illustration of what happens with the spherical-gain rule, let us assume that the total points one uses equals 10 (any other number would do as well). With the spherical-gain rule, one's best strategy is to bet strictly in accordance with one's true beliefs. If one put all one's bet on the alternative one considered to be most likely correct, the one with probability .7 in our example, the expected gain would be

$$\frac{7}{10} \text{x} \left(\frac{10}{\sqrt{10^2 + 0^2}} \right) + \frac{3}{10} \text{x} \left(\frac{0}{\sqrt{10^2 + 0^2}} \right) = .70,$$

whereas if one bet in accordance with one's true beliefs, placing .7 of one's bet on the .7 alternative and .3 of it on the other one, one's expectation would be

$$\frac{7}{10} \text{x} \left(\frac{7}{\sqrt{7^2 + 3^2}} \right) + \frac{3}{10} \text{x} \left(\frac{3}{\sqrt{7^2 + 3^2}} \right) = .76.$$

The superiority of scoring rules that have the matching property over those that do not have it is indisputable from a mathematical point of view; however, the relevance of this distinction to much of human probabilistic reasoning is debatable. The evidence that people behave very differently when scoring rules are admissible than when they are not is thin, and in many, if not most, of the real-life situations in which probabilistic reasoning is required, outcomes, which are controlled by nature, do not demonstrably have the matching property in any case. The use of such scoring rules for test administration is an interesting possibility, however, especially when tests could be administered under computer control so that test takers could get immediate feedback that might help them develop a good understanding of the advantages of honestly reflecting their true belief states in their answers. Admissible scoring rules have other characteristics that could be exploited to advantage in providing much finer-grained feedback to teachers regarding the knowledge states of students and classes in the aggregate than do the results of conventionally administered multiple-choice tests.

The Illusion of Control

One way to run a lottery is to permit players to purchase tickets with preassigned numbers. Another is to permit the ticket buyers to specify the numbers they wish to play. There is some evidence that players find the second approach more to their liking than the first. In one study, in which a lottery was conducted by drawing names of professional football players from a bag, participants who selected the name of a player valued their tickets about four times as much as did participants who were assigned a name at random (Langer, 1975). Langer's explanation of this difference was that the ability to choose gives one an *illusion of control.*

Apparently people believe—or at least act as though they believe—that their chances of winning are greater when they select an item that is to be part of a random draw than when they are assigned one. We know that many people who play lotteries do not choose random numbers, but resort to a variety of methods for selecting what they hope to be lucky ones. Given Langer's (1975) findings, we would have to judge as highly rational the behavior of lottery operators who have opted to permit ticket buyers to select their own numbers, because such lotteries should entice more players than those in which the player has to take whatever happens to be on the next available ticket.

Paulos (1998) describes an interesting way in which feedback from lotteries can provide information that can be interpreted, erroneously, as evidence that players stand a better chance of winning if they pick their own numbers than if they accept numbers picked by machine. Consider a lottery that has only two players, and the winner is selected by a random drawing of a number between 1 and 10. Suppose that player A always plays her favorite number, say 7, whereas B plays a different number each time. Over many games, A and B will win with about equal frequency, but the number 7 will be the winning number more often than any other number; 7 is the winning number on *all* of the 50% of the games that A wins, whereas on the other 50% of the games, which B wins, the winning numbers are distributed over the entire range, so no one number is winner for B on more than a small percentage of the games. If, in a large lottery, many of the people who pick their own numbers stay with the same number week after week, whereas machine-picked numbers vary randomly over all the possibilities, it can happen that hand-picked *numbers* have a higher probability of winning than do machine-picked numbers, whereas *individual players* who hand pick their numbers do not have a higher probability of winning than do those who buy numbers picked by machine. Paulos points out that, in an analogous way, it can happen that "any situation having many outcomes, one of which is conventionally favored by the society at large, will seem to generate the conventionally favored outcome more frequently than chance would suggest" (p. 184).

Some people, including, perhaps especially, those who gamble, make a distinction between chance and luck (Keren & Wagenaar, 1985; Wagenaar & Keren, 1988; Wagenaar, Keren, & Pleit-Kuiper, 1984). Luck is seen as a causal agent that sometimes overrules the effects of chance. Good luck, for example, might be credited with an unexpected run of wins in a game of chance, such as roulette. A run of losses is attributed to bad luck. Belief in luck in this context can be reinforced by the tendency to underestimate the probability of relatively low-probability chance events, such as several successive tosses of heads on a coin. Inasmuch as the probability of alternations is typically overestimated whereas that of repetitions is underestimated (Budescu, 1985; Falk, 1981; Falk & Konold, 1997; Kubovy & Gilden, 1990; Lopes & Oden, 1987; Am. Rapoport & Budescu, 1992), a run demands an explanation as a non-chance event, and the idea of luck satisfies that demand. Belief in luck can give one a false sense of control over the situation: It supports the assumption that continuing to play when luck is with one, and quitting when it is not, improves one's chance of winning.

Falk (1991) points out that beliefs in the "hot-hand" phenomenon and in luck are analogous in interesting ways:

> You only have to exchange these two terms with each other and you get the same story. People believe that one cannot force luck to happen. One should wait till luck appears and know how to utilize it wisely. By the same token, a hot hand cannot be summoned at will, but once a player is "hot" it is important for players on the team to pass the ball to that player (and for the opposition to watch him closely). In both contexts, people are confronted with random sequences of wins and losses and their attention is drawn to the subjectively overlong runs. Far from adjusting their concept of what could happen by chance, they now invoke an interpretation in the form of "luck" or "hot hand." The gambler's fallacy in perception of randomness thus appears in different disguises and surely it does so in many other contexts. (p. 217)

The illusion of control has been hypothesized to be responsible to some degree for the fact that people often tend to be overconfident of their predictions about future events over which they believe themselves to have some degree of control (Griffin et al., 1990; S. J. Hoch, 1985); attorneys, for example, are likely to predict trial outcomes that are better than those they actually get (E. F. Loftus & Wagenaar, 1988; Wagenaar & Keren, 1986), and people generally appear to be more optimistic about events that are perceived to be under their control than about those that are not (Budescu & Bruderman, 1995; DeJoy, 1989; P. Harris, 1996; Hoorens & Buunk, 1993; Zakay, 1984). The idea is that the degree of control that people have over the anticipated events or performance typically is less than they think they have and their confidence in their predictions is based on what they think they have (Alloy & Abramson, 1979; Bradley, 1981; Dickinson, Shanks, & Evenden, 1984).

Closely related to the idea of illusion of control is the question of the willingness or reluctance to take responsibility for uncertain events. Can anyone doubt that most of us are more willing to take credit for decisions that turned out well than for those that turned out poorly, independently of the quality of the decisions as judged strictly in terms of the information available at the time they were made? Good decisions can, of course, turn out poorly and poor decisions can turn out well. Nevertheless, decision makers are usually rewarded positively or negatively on the basis of the outcomes of the decisions they have made. People who make poor decisions that, because of unanticipatable events, turn out to have desirable consequences are likely to be seen as astute decision makers and to reap rewards that are commensurate with the desirability of the outcomes. Similarly, people who make excellent decisions that, again because of unforeseeable circumstances, turn out to have undesirable effects, are likely to feel fortunate if the worst personal consequence is lack of recognition for the high quality decisions they made.

The Force of Personal Experience

People are much influenced in their probabilistic thinking by vivid personal accounts of individual concrete cases, often more than by dry summaries of abstract incidence statistics, even though the latter may be a more reliable basis for prediction than the former. One possible reason for the force of personal experiences is that we remember them better than things we have learned about secondhand. If we remember them better, they are likely to be more available than other information for reasoning tasks for which they are relevant. It is also possible that we have more confidence in personal experiences than in information we acquire in other ways; we know the experiences occurred whereas believing what we read or hear requires an element of faith. The tendency to generalize from very small samples can be seen as illustrating the force of personal experience, because the small samples from which such generalizations are made often are samples that one has experienced directly, and consequently they are easier to call to mind than the larger samples that one *might* have experienced but did not.

SUMMARY

Choosing—selecting among alternatives—is a ubiquitous human activity. Choices frequently must be made with uncertain knowledge of what their consequences will be. The idea of making choices under uncertainty in such a way as to maximize expected utility has been central to normative theories of decision making for a long time—the distinction between utility and monetary value having been recognized at least since the 18th century. However, much research has shown that people often make choices in ways that appear not to be dictated by the intention to maximize expected utility. Recognition of this fact has motivated efforts to construct theories or models of choice behavior that are descriptive of the choices people actually make under conditions of uncertainty.

Well-known theoretical treatments of choice behavior that are intended to be descriptive of how people actually behave include Simon's (1955, 1957) theory of bounded rationality, Kahneman and Tversky's (1979b) prospect theory, and Gigerenzer's (1991b, 1993) theory of probabilistic mental models. Much recent research, pioneered by Kahneman and Tversky (1973; Tversky & Kahneman, 1974), has focused on the identification and assessment of a variety of heuristic strategies that people use in making choices under uncertainty. Evidence indicates that although the use of strategies that simplify choices often yields undesired outcomes, it also can sometimes be very effective in providing satisfactory solutions to complex problems with limited cognitive effort (Gigerenzer & D. G. Goldstein, 1996; Gigerenzer & Todd, 1999a).

Research has focused also on the identification of specific variables that can affect choice behavior for better or worse, and especially the latter. The discovery of the importance of *framing effects*—that preferences among decision outcomes are determined not only by what the outcomes are, but by the way in which they are represented—is illustrative of this work. Numerous other factors that can influence choice under uncertainty have been identified and experimentation continues on what has proved to be an exceptionally interesting and fruitful line of inquiry.

CHAPTER 11

People as Intuitive Probabilists

We develop subjective concepts of probability which permeate and guide our thoughts and actions.

—Cohen (1957, p. 128)

Applying probabilities and statistics is much more a matter of grasping the situation, constructing informal arguments, and building comprehensive narratives than of substituting numbers into formulas.

—Paulos (1998, p. 82)

In general, we cannot expect good quantitative statistical intuitions, nor even good qualitative intuitions, for probability questions of a sort that do not arise in ordinary experience. But we would expect good intuitions to the extent that pragmatic conditions in the world would provide the required tuning to experience.

—Margolis (1987, p. 164)

PEOPLE AS BAYESIAN REASONERS

Much of the research that falls under the general rubric of intuitive statistics has had to do with the ways in which people process probabilistic data. How are beliefs about the world formed or modified as a consequence of the receipt of data that are relevant to those beliefs but insufficient to demonstrate their truth or falsity conclusively? More specifically, research has focused on how data of a probabilistic nature are used to modify beliefs in situations for which Bayes's rule would be an applicable belief-revision tool.

It is of more than passing interest that two of the more extensive lines of investigation appear to have led to opposite conclusions regarding a fundamental characteristic of human beings as processors of probabilistic information. On the one hand is a body of research that suggests that people are overly conservative in their use of probabilistic information and tend not to revise existing beliefs as much as they should in the light of newly acquired data. On the other hand are numerous experiments that support the idea that people are relatively insensitive to "base-rate" information when making probability judgments and tend to base those judgments almost entirely on newly acquired "case-specific" data. In what follows, I shall first describe representative findings from both of these lines of research and then consider the question of how to reconcile them.

Conservatism in Probabilistic Judgments

One of the earliest and best documented findings regarding how well people do, relative to what Bayes's rule says they should do, in extracting information from data is that when people estimate posterior probabilities $p(H \mid D)$—when they attempt to revise estimates of the probability that some particular hypothesis is true upon receiving data that relate to that hypothesis—they tend to revise their previous estimates in the appropriate direction but not by sufficiently large amounts (Donmell & DuCharme, 1975; Edwards, 1968; Edwards, Lindman, & Phillips, 1965; Messick & Campos, 1972; Navon, 1978; C. R. Peterson & Beach, 1967; C. R. Peterson & DuCharme, 1967; C. R. Peterson & A. J. Miller, 1965; C. R. Peterson, Schneider, & A. J. Miller, 1965; Phillips & Edwards, 1966; Phillips, Hays, & Edwards, 1966; Am. Rapoport & Wallsten, 1972; Slovic & Lichtenstein, 1971). This result has been described as evidence of people's conservatism as Bayesian information processors.

People tend, according to this view, to extract less information from data than is there to be extracted. Another way to say this is that people require more evidence than does an ideal Bayesian process to arrive at a given level of certainty regarding which of several competing hypotheses is true. If one begins by assuming that each of several possibilities is equally probable, the effect of this type of conservatism is to keep the estimator at all times closer to the condition of maximum uncertainty (equally likely alternatives) than is appropriate, which means that at any given time the estimate of a posterior probability will be too low for high probabilities and too high for low ones, relative to what Bayes's rule would produce.

Conservatism in the estimation of posterior probabilities has been described as the primary finding of Bayesian research (Slovic et al., 1977). It was this finding that prompted Edwards (1965) and others to propose that probabilistic information-processing systems should use human experts to provide estimates of data conditional upon hypotheses, $p(D \mid H)$, for each of the hypotheses under consideration and to have machines generate posterior probabilities, $p(H \mid D)$ in accordance with Bayes's rule, using the conditional probabilities provided by the experts as inputs.

Why should conditional probabilities be easier to estimate accurately than posterior probabilities? One possibility is that it is easier to think in terms of contingencies in the direction of *D* given *H* (or *H*, therefore *D*) than in that of *H* given *D*, because we find it more natural to reason from cause to effect. It may be more natural in medicine, for example, to think in terms of the probability that one has a particular symptom if one has a particular disease, because the symptom is perceived as a consequence of the disease and not the reverse.

In addition, formal training programs may reinforce any tendency to think in the *D*-given-*H*, or if-*H*-then-*D*, direction. Consider medicine again. Medical students are taught what to expect by way of symptoms, given a specified disease. Less often are they encouraged to think in terms of the various diseases for which a particular symptom could be a sign. One suspects this may be true of teaching also in other areas, say electronic or automotive troubleshooting, or criminal investigation.

Though most of the studies of posterior probability estimation have yielded evidence of conservatism and have been consistent with the idea that human estimators are better at estimating $p(D \mid H)$ than $p(H \mid D)$, there have been some exceptions. Occasionally when the hypothesis set has been small and the environment frequentistic, people have produced posterior probability estimates that were more extreme than those produced by the application of Bayes's rule (Southard, Schumn, & Briggs, 1964). Messick and Campos (1972) found that conservatism can be diminished by encouraging people to think of probability as reflecting the ratio of favorable-to-total possible outcomes rather than strength of belief. Other studies that cast doubt on the generality of the finding of conservatism in the estimation of posterior probabilities include Howell (1966), R. J. Kaplan and Newman (1966), and Schum, I. L. Goldstein, and Southard (1966).

Conservatism, as represented by the results of the studies that have found it, could be the consequence of an overweighting of the importance of prior probabilities or, what amounts to the same thing, from an insensitivity to the diagnosticity of incoming sample data; although it appears not to be the case that people are totally insensitive to diagnosticity (Devine, Hirt, & Gehrke, 1990; Skov & Sherman, 1986; Snapper & C. R. Peterson, 1971; Trope &

Bassock, 1982, 1983). It could be due to inappropriate application of Bayes's rule. And it could stem in part from a (not necessarily unreasonable) unwillingness of people always to assume that the data selections are determined completely by chance (L. J. Cohen, 1982).

Another possibility is that people discount evidence to compensate for the fact that it often is unreliable in real-world situations. One might argue that people should not discount evidence in laboratory situations in which the data are unambiguous and completely trustworthy, but it is conceivable that they do so as a general rule and that that habit serves them well in most of the cases they encounter in the everyday world. The effect of the various normative approaches that have been developed to deal with data that are known, or assumed, to be unreliable is to discount evidence, thus moving probability updates in the direction of conservatism. When people have been asked to make updates on the basis of data of different specified degrees of reliability, they have tended to be overly conservative with highly reliable data, and to become less conservative as reliability is decreased, sometimes approaching optimality for intermediate degrees of reliability, and becoming insufficiently conservative for especially low ones (E. M. Johnson, 1974; E. M. Johnson, Cavanagh, Spooner, & Samet, 1973; Schum et al., 1971).

Which of the various possibilities, or combination thereof, accounts for the phenomenon is not clear, despite the considerable attention that has been given to this question (Fischhoff & Beyth-Marom, 1983). Almost from the beginning of work on Bayesian decision making, there have been proponents of the view that conservativism results from miscalculation of conditional probabilities as well as for the alternative view that people can calculate, or estimate, the conditional probabilities well enough, but they misapply the theorem as a whole. To the extent that one attributes the finding of conservatism to computational difficulties, as distinct from the overweighting of prior probabilities or the underweighting of updating information, one may question whether conservatism is quite the right concept to apply.

Base Rates

A prior probability distribution, in the Bayesian approach to statistical reasoning, is a representation of what the situation was believed to be before a specific piece of information became available; it means prior with respect to that piece of information. Similarly, posterior means after that particular information became available. The difference between the prior and posterior distributions reflects the impact the piece of information in question had on the Bayesian assessment of the situation.

It should be clear from the computation involved in producing a posterior distribution that the probabilities that summarize what one knows about a situation before receiving a new item of information remain relevant after the further specific information is received. Bayes's rule is a prescription for combining the new information with the old so as to update one's probabilistic assessment of the situation; this prescription does not disregard the old information but rather uses it as the point of departure for establishing a revised view. The same bit of new data will yield different posterior distributions given different prior distributions.

Prior probabilities and base rates are not conceptually identical—probability, as we have seen, can have more than one connotation, but base rate is clearly a frequentistic concept, which refers to the frequency, or relative frequency, of some event as an actuarial fact. Many theorists hold that prior probabilities should reflect base rates when the latter are known. Thus, if it is known that 80% of the taxicabs in some town are green, the prior probability that a randomly encountered cab will be green should be considered .8, according to this view. Of course, base rates often are not known by individuals who find themselves in situations where knowledge of them would be useful, and when that is the case, they can hardly be faulted for not using them. Laboratory studies of base-rate neglect typically have provided the relevant base rates, either directly or indirectly; base-rate knowledge regarding real-world events has been seen as an important aspect of what it means to be an expert (J. F. Yates, 1982).

People Often Ignore or Discount Base Rates

Numerous experiments have yielded results that suggest that, upon receipt of specific information about a probabilistic situation, people tend to rely on it alone and to ignore, or underweight, prior probabilities, as represented by base rates, more or less completely (Bar-Hillel, 1980; Bar-Hillel & Fischhoff, 1981; Borgida & Brekke, 1981; J. J. Christensen-Szalanski & Beach, 1982; Grether, 1980; Lyon & Slovic, 1975, 1976). Contrary to the Bayesian prescription, the new information seems to supplant the old entirely, or nearly so. Failure to give due weight to base rates has been viewed as a particularly well documented and pervasive manifestation of the shortcomings of people as intuitive statisticians.

On the face of it, the results of many experiments seem to support the idea that people do indeed give inadequate weight to base rates in their estimates of posterior probabilities. There is some question as to whether participants in experiments always understand what the probabilities used in these problems mean and whether they may confuse, say, the probability of a witness's report conditional on the actual color of a cab with the probability that a cab was a given color conditional on the witness reporting that color (Baron, 1988). This is a worry and dictates caution in the interpretation of results. On the other hand, the finding that people tend to ignore, or severely discount, base rates in such problems is a robust one. It has been obtained with a variety of problems and has proved to persist despite efforts by experimenters to increase the saliency of the base rates in several ways.

Hammerton (1973), for example, used a disease-detection problem like those discussed in chapt. 7 in which the following information was pertinent. A particular diagnostic test will give a positive result with a probability of .9 when applied to a person who has the disease; when applied to a person who does not have the disease, the same test will give a positive result with a probability of .01. About 1% of the population has the disease. Mr. Smith has been tested and the test result shows positive. What is the probability that Mr. Smith has the disease?

Letting $p(H \mid D)$ and $p(D \mid H)$ represent, respectively, the probability of having the disease (Hypothesis) given a positive test result (Data) and the probability of getting a positive test result given one has the disease, Bayes's computation is

$$p(H|D)=\frac{p(D|H)p(H)}{p(D|H)p(H)+p(D|{\sim}H)p({\sim}H)},$$

or

$$p(H \mid D) = (.9)(.01)/[(.9)(.01) + (.01)(.99)] = .476.$$

Hammerton's participants judged the probability to be closer to .9, apparently giving far more weight to the results of the test than to the population base rates.

Of course use of .01 as the prior probability that Smith has the disease (and .99 as the prior probability that he does not) is justified only on the assumption that Smith was randomly selected for testing—that, in other words, the prior probability that he has the disease is accurately reflected by the incidence of the disease in the general population. The plausibility of this assumption is dubious at best in most real-life situations; people usually get tested for a specific disease because there is some reason, other than the fact that the disease exists within the general population to which they belong, to suspect that they may have it.

One cannot rule out the possibility that people apply their understanding of this fact in the experimental situation. If it were possible to determine for diseases with known incidences of occurrence in the general population what proportion of the individuals who get tested for those diseases have them, it would be surprising if the latter proportions were not considerably greater than the former. This raises the following question. Suppose one has a diagnostic test for a particular disease, which turns out to be positive, and one wishes to calculate in a Bayesian fashion the posterior probability that one has the disease.

What should be used as the appropriate base-rate or prior probability? The rate of incidence of the suspected disease in the general population? Its rate among people who, for whatever reason, have taken the test? Its rate among people who take the test for reasons similar to one's own? One's own pretest estimate of the probability of having the disease? Something else?

Complications of this sort can often be identified in experimental problems involving courtroom, medical, or other situations of which people have certain knowledge or expectancies in addition to what they obtain from the experimental protocol. The following problem, used by Lyon and Slovic (1976), seems to avoid these complications:

> A light bulb factory uses a scanning device which is supposed to put a mark on each defective bulb it spots in the assembly line. Eighty-five percent of the light bulbs on the line are OK; the remaining 15% are defective. The scanning device is known to be accurate in 80% of the decisions, regardless of whether the bulb is actually OK or actually defective. That is, when a bulb is good, the scanner correctly identifies it as good 80% of the time. When a bulb is defective, the scanner correctly marks it as defective 80% of the time. Suppose someone selects one of the light bulbs from the line at random and gives it to the scanner. The scanner marks this bulb as defective. What do you think is the probability (expressed as a percentage) that this bulb is really defective? (p. 292)

It is hard to think of knowledge or expectancies obtained outside the experimental situation that would get in the way of a clear understanding of this problem or discount the relevance of the base rate of defective bulbs. Explicit mention in the statement of the problem of random selection from the population should call attention to the base rate and ensure its salience. Nevertheless, the results obtained with this problem were essentially identical to those obtained with the cab and disease-diagnosis problems. For some participants, Lyon and Slovic (1976) used the light bulb problem with 80% scanner accuracy (as in the previous problem statement); for others they used 50% and for still others 20%. In all cases, the median estimate produced by the participants matched the specified accuracy of the scanning device independently of the base rate. And performance was not affected when the base rate of defective bulbs was changed from 15% to 1%. These results provide strong evidence that people are indeed relatively insensitive to base rates when they have conditional-probability data in hand.

Numerous other studies have corroborated the conclusion that people tend to ignore or to discount base rates when case-specific information is available (Cassells, Schoenberger, & Grayboys, 1978; Eddy, 1982; Hammerton, 1973; Landman & Manis, 1983; Meehl & Rosen, 1955; Pollard & Evans, 1983). This has proved to be so even when the case-specific information provided has been designed to be totally irrelevant to the task at hand (Kahneman & Tversky, 1973). Nisbett, Borgida, Crandall, and Reed (1976) give several examples of the power of case studies to persuade, and of the impotence of base rates against them. Apparently many people do not believe that base rates are relevant to probability judgments, or at least not universally so.

There is some evidence that people give insufficient weight both to population base rates and to what might be considered personal base rates (frequencies of past personal experiences of specific types) when predicting future personal behavior (Buehler, Griffin, & Ross, 1994; Kahneman & Tversky, 1973; Nisbett & Borgida, 1975; Osberg & Shrauger, 1986). This may help to account for the fact that we find it so easy to underestimate the time it will take us to do specific tasks, despite the experience of having frequently underestimated in the past (Buehler, Griffin, & Ross, 1994; Hayes-Roth & Hayes-Roth, 1979; Kidd, 1970).

The neglect or discounting of base rates has often been taken as evidence of irrationality, at least insofar as rationality is represented by probability theory, and Bayes's rule more specifically. This interpretation can be challenged, however, even granting that base rates are indeed neglected, at least under some conditions. One problem with it, pointed out by Gigerenzer (1991b, Footnote 2) is that the terms "base rate" and "prior probabilities" are often used interchangeably in the psychological literature, but, as already noted, they are not synonymous. Prior probabilities are what are expressed in Bayes's rule; base rates, among other considerations, may be used as the basis for estimating these probabilities. In the absence of any other relevant knowledge of the situation, one might make the prior probabilities equal to the base rates, but often other knowledge is available and when that is the case it should not be ignored. Macdonald (1997a) points out that naturally occurring events can be classified in many ways and, because different subsets having different probabilities can be identified, randomness may be impossible to establish. Moreover, because randomness does not have a single universally agreed upon meaning, "even when it is stated that an event has been randomly sampled, randomly could be interpreted as haphazardly and the event could still be described as having been selected from any number of populations each with a different base rate" (p. 778)

But Not Always

Thus the evidence is strong that people often discount or ignore base rates. I want to turn now to some evidence that people do not always do so, but before doing that we should note a methodological problem that complicates the interpretation of some studies of base-rate use. Sometimes it is not possible to determine from data whether participants in a Bayesian information-processing experiment were behaving in accordance with Bayes's rule or not, because it is not clear what assumptions they made to fill in essential information missing from problem descriptions. A critical assumption underlying the application of Bayes's rule is that each observation is randomly and independently selected from the entire set of hypothesized possibilities. When what constitutes that set is not clear, the base rates will be ambiguous, and in order to judge the rationality of people's estimates, one must know what assumptions they have made to resolve that ambiguity.

But leaving that problem aside, let us accept as fact that people often fail to pay much attention to base rates when reasoning about uncertain events. The point to be made now is that they do not always do so. There are certain conditions under which base rates are utilized and it appears to be possible to increase the chance that they will be taken into account in other cases as well. Base rates are used, for example, when case-specific information is not provided, which is to say when the base-rate information is all one has (Kahneman & Tversky, 1973). It is important to bear in mind, however, that even irrelevant case-specific information can swamp base-rate information, in some circumstances; apparently having *no* case-specific information and having *some* are quite different situations, psychologically, independently of how useful or useless the case-specific information may be.

When case-specific information is available, base-rate information may be used more appropriately if it is obtained from experience in real-world situations than if it is provided verbally in laboratory situations (J. J. Christensen-Szalanski & Bushyhead, 1981). It has been suggested that, in general, one is more likely to use appropriate statistical reasoning when the statistical characteristics of the situation are objective and relatively apparent than when they are more subjective or obscure (Jepson et al., 1983; Nisbett et al., 1976).

Even in laboratory situations base rates may be used if the information that is critical to their relevance is sufficiently salient. For example, in one experimental paradigm that has been used, a person is told the relative sizes of profession A and profession B, given some biographical information about an individual, and asked to estimate for each of the professions the probability that the individual is a member of it. The usual finding is that people pay less attention to the information regarding the relative sizes of the professions—the base-rate information—than they should, according to Bayes's rule, and more to the biographical, or case-specific, information. It is important to note that the base-rate information is relevant here only on the assumption that the individual whose profession is being judged was randomly selected from the population that includes the combined sets of the two professions. Gigerenzer, Hell, and Blank (1988) have shown that when the validity of this assumption is made salient by having people draw the biographical descriptions from an urn in which the two professions are represented in the specified proportions, the base-rate information is not ignored.

Other experimenters have been able to increase people's use of base rates by making their random sampling from a population salient (R. D. Hansen & Donoghue, 1977; Wells & Harvey, 1977) or by having them experience the statistics of a situation by witnessing a series of incidents, one at a time (J. J. Christensen-Szalanski & Beach, 1982; Schlotterbek,1992). Fischhoff, Slovic, and Lichtenstein (1979) found they could improve the use of base-rate information by giving the same participants several versions of a word problem that were identical, excepting for base rates; presumably the saliency of the base-rate information was assured by this design.

There is also evidence suggesting that base rates are likely to be used when they are expressed in sufficiently concrete terms (Manis, Dovalina, Avis, & Cardoze, 1980), when they are expressed as frequencies (50 out of 1,000) rather than as probabilities or percentages (.05 or 5%) (Cosmides & Tooby, 1990; Gigerenzer, 1993, 1994; Gigerenzer & Hoffrage, 1995), when their relevance to a particular question is made clear (Brekke & Borgida, 1988), and especially if there is an apparent causal relationship between them and the cases of interest (Ajzen, 1977; Bar-Hillel, 1980; Tversky & Kahneman, 1978). When the taxicab problem is recast so that .85 and .15 represent the proportion of accidents involving blue and green cabs respectively, rather than the proportion of blue and green cabs on the road, the base rates are more likely to be taken into account (Bar-Hillel, 1980, 1983, 1984). When the statement of the taxicab problem provides the information that 85% of the cabs on the road are blue and 15% are green, use of these numbers as base rates requires the assumption that the probability that the cab involved in the accident in question was blue (green) is exactly the same as the proportion of all city cabs that are blue (green), which is tantamount to assuming that the probability that any given cab will be involved in an accident is independent of its color. This seems a reasonable default assumption in the absence of information to the contrary, but it is an assumption. Bar-Hillel's result suggests that people may fail to make this assumption in the absence of explicit instructions to do so.

Although the data clearly demonstrate that people may become more sensitive to base rates if they explicitly engage in random sampling, or perhaps are helped in some other way to think in frequentistic terms, they do not support the conclusion that base-rate neglect can always be eliminated in this way (Camerer, 1990; Griffin & Tversky, 1992). Gigerenzer (1991b) also cautions that simply telling people that cases were randomly sampled may not suffice to convince them of that, especially in instances in which random sampling would not be expected in the real-world situations the experimental scenarios are intended to represent.

Base rates have been taken strongly into account also when problems that are isomorphic with some that have shown base-rate neglect are explicitly framed as games of chance (Ginossar & Trope, 1987). It appears too that base rates can be made more salient through training people to pay attention to them (J. J. Christensen-Szalanski & Beach, 1982), or by motivating them to do so by, for example, having them—in the blue–green cab scenario—play the role of a lawyer for one of the cab companies (Ginossar & Trope, 1987).

Gavanski and Hui (1992) found that people showed a high degree of sensitivity to variations in base rates when the relevant sample spaces for judgments corresponded to natural-kind categories, but not when the relevant sample spaces were not natural-kind categories. Natural-kind categories are categories in terms of which people spontaneously partition the world and are to be distinguished from categories that may be defined for special purposes but are not commonly used by people in their everyday experience. Gluck and Bower (1988) had people classify people as having either a rare or common disease on the basis of consideration of a set of symptoms; classifications were sensitive both to the rarity-commonality of the disease (base rates) and to the relative likelihoods (likelihood ratio) of the symptoms given the disease.

Results from another study suggest that base rates are used in making estimates of covariation significantly more if they relate to causal factors than if they relate to noncausal factors, even though equally valid predictions can be made in both cases (Ajzen, 1977). People who were told that students who spend more than 20 hours per week studying have higher grade-point averages than students who spend less than 20 hours used that base-rate information in evaluating information about cases to predict grade-point averages; in contrast, people who were told that students who lived more than 20 miles from campus had better grades than those who lived closer placed less weight on information about distance of commute, even though it presumably could have been as useful in making predictions. A causal connection between study time and grade-point average is easy to imagine, whereas one involving distance of home from campus is not so easy to see.

People also show some intuitive sensitivity to base rates in comments and attitudes about low-probability events. Nisbett et al. (1983) point out, for example, that the statement "I can't understand it; I have nine grandchildren and all of them are boys" would not be perceived as strange by most people, but substitute three for nine in that sentence and it might be. The student who fails a test is likely to be less upset if three quarters of the class failed it than if nearly everyone passed it. But granted that people sometimes show a sensitivity to the relevance of base rates, they often do not, and it is this fact that many investigators feel requires explanation.

A practical problem in applying base rates in everyday reasoning is that of figuring out what they are. Often it is very difficult, if not impossible, to do this, or even to determine what would constitute a reasonable assumption in that regard. Even if one can identify the populations the relative sizes of which are the critical data, the relative sizes may not be known and what to assume about them may not be clear. When relevant numbers are available, they may be ambiguous. As we have already noted, for example, learning that a particular test that is used for medical disgnosis is 90% accurate does not suffice, because 90% accuracy can have a variety of meanings not all of which imply the same base rates. Moreover, one cannot always take reported base rates at face value. It is easy to find examples of grossly inaccurate reports of base rates in public media and, as J. R. Anderson (1990) points out, this fact might account, to some degree, for the tendency of people in psychological experiments to take those that they encounter in that context less seriously than they might.

Theorists do not agree regarding exactly how base rates should be taken into account in estimating prior probabilities. Many appear to assume that prior probabilities should be equated with base rates when the latter are known. But others do not make this assumption (Levi, 1983; Okasha, 2000). Okasha makes the important point that whether or not one uses base rates for prior probabilities is not, strictly speaking, a Bayesian issue. Bayes's rule has to do with how one should use data to modify prior probabilities; it does not specify how the prior probabilities are to be determined.

Reconciling Conservatism and Base-Rate Discounting

At first glance, conservatism and base-rate discounting—the ideas that people generally give insufficient weight to newly acquired data in revising preexisting probability estimates and that they usually ignore or take too little account of base rates—appear to be mutually contradictory. If we knew only that they were guilty of the one crime, we would expect them to be innocent of the other. Indeed, one of the earliest explanations proposed for the finding of conservatism in Bayesian decision making was that people put too much emphasis on base rates and consequently do not extract from newly acquired data all the information they contain (Edwards, 1968).

Gigerenzer et al. (1989) point out this apparent inconsistency in the results of two lines of experimentation and characterize it as something of an embarrassment to psychological research: "The question of why people seemed to be conservative during the 1960s and anti-conservative after 1970 has not yet been answered, if only because it was almost never posed. It should be very disturbing that established facts suddenly do an about face. But the new facts were instead enthusiastically received as revelatory of underlying mental heuristics, and the opposite facts largely ignored as too old to be true" (p. 219). Fischhoff and Beyth-Marom (1983) note that the line of research on conservatism was "quietly abandoned" without establishing the roles played by the several factors that had been hypothesized to account for it, and attribute the cessation of activity on this question partly to the discovery of base-rate neglect, which they see as the antithesis of conservatism.

In any case, the finding of base-rate neglect seems to be inconsistent with the idea, which has considerable experimental support, that people often are very reluctant to modify beliefs they hold even when they receive quite strong evidence that modification is justified (Nickerson, 1998). However, there is a difference between the typical experimental situations in which conservatism and base-rate discounting have been studied that may help to show the inconsistency to be in appearance only. In the former case, people usually are given a distribution of probabilities, or something more or less equivalent to this, that is claimed to represent what is known about the situation before any data are obtained. Assuming they accept this distribution for what it is represented to be, the task then is to revise an existing belief state in the light of the data that are subsequently acquired. In the typical base-rate study, one is given base-rate information and case-specific information at the same time, and it may not always be apparent to the participant that the base-rate information represents what one should believe about the situation before processing any case-specific data. In other words, it may be less clear in this instance that the case-specific data are to be used to revise a preexisting belief state.

The difference is sharpened considerably when we compare the conservatism that relates to the modification of meaningful beliefs that people actually hold with the neglect of base rates as that phenomenon is observed in laboratory situations. Base rates typically have been supplied by the experimenter and have involved variables of no special interest to participants beyond their relevance to the experiment. It perhaps should not surprise us to learn that people's behavior with respect to experimenter-provided base rates that have little to do with meaningful beliefs that they hold does not give an entirely unequivocal indication of how they are likely to treat data that relate to personal beliefs that they have formed and possibly held for a considerable time.

Consensus Information

The failure of people to give due weight to base-rate information when making judgments about category membership has a parallel in the way people tend to discount consensus information when ascribing causes to human behavior. In this case when asked to predict the behavior of individuals in specified situations, people sometimes tend to ignore available information regarding typical behavior of people in those situations in the past. The finding is illustrated by an experiment by Nisbett and Borgida (1975).

These investigators informed people of the results of two earlier experiments, one by Nisbett and Schachter (1966) in which participants were asked to take as much electric shock as they could stand, and another by Darley and Latane (1968) in which participants witnessed what appeared to be a fellow participant in an adjacent experimental cubicle having a seizure. Nisbett and Borgida's subjects were also shown taped interviews with individual participants in the earlier experiments or were shown written descriptions of their backgrounds and personalities and were asked to guess how the individuals behaved in the earlier experimental situations. Some of Nisbett and Borgida's subjects were informed of the actual behavior of the participants in the earlier experiment, *as a group*. The predictions of individual behavior were not strongly influenced by the information about their aggregate behavior; those of the subjects who had that information were similar to those of the subjects who did not have it, and in neither case did they match the actual behavior very closely. Also, the knowledge that most participants in the Darley and Latane experiment had not offered to help the person who appeared to be having a seizure did not make Nisbett and Borgida's subjects less likely to blame the individual participants in the earlier experiment for their failure to help.

Nisbett et al. (1976) equate consensus information with base-rate information, the difference being that consensus information is base-rate information about behavior, rather than about category membership. Attribution they see, however, as a more complicated and less direct inference than a prediction. They contrast the difference between the two inferential processes this way:

> Kahneman and Tversky ask their subjects to produce a rather direct and uncomplicated chain of inference: "If the majority of the members of the population belong to a particular category, then odds are the target case does also." Their subjects fail to make such an inference. In the attribution research we have been discussing, a still more elaborate chain of inference is requested: "If the majority of the members of a population behave in a particular way, then the situation must exert strong pressures toward that behavior, and therefore it is unparsimonious to invoke personal idiosyncrasies to account for the behavior of the target case if his behavior is modal." (p. 124)

Inasmuch as there is evidence, which Nisbett et al. review, that people fail to apply behavioral base rates to predictions about target cases, the question of why people ignore consensus information in making attributions reduces, in their view, to the question of why they disregard base-rate information in general. The topic of attribution—especially in the form of the question of how people attribute effects to causes—has received a great deal of attention from researchers; the role of base rates is but one aspect of this interest.

Gigerenzer (1994) makes the thought-provoking observation that although psychologists have tended to take the neglect of base rates by participants in psychological experiments as evidence of irrationality, experimenters themselves typically have ignored base rates in their own testing of hypotheses; and have not seen the inconsistency in their behavior in this regard. A large majority of statistical tests that are reported in the psychological literature—that which deals with base-rate neglect included—are significance tests of the Fisherian type, which do not take base rates into account.

Primacy Effects

Intuitively we would expect that, given *n* items of information that are relevant to some conclusion that is to be drawn or decision that must be made, a rational process for integrating those items of information should be insensitive to the order in which they are considered. Bayes's rule has this property: *n* observations that are used to update a distribution of probabilities over a hypothesis set will have the same combined effect independently of the order in which the observations are made.

The conclusions that people draw and the decisions they make often are not insensitive to the order in which the data that helped determine those conclusions or decisions were obtained (Pennington & Hastie, 1988; Walker, Thibaut, & Andreoli, 1972). In particular, information that is acquired early in the process often is given more weight than that acquired later. C. R. Peterson and DuCharme (1967), for example, had people sample a sequence of colored chips and estimate the probability that the sequence came from an urn with a specified distribution of colors rather than from a second urn with a different distribution. The sampling was arranged so that the first 30 trials favored one urn and the second 30 favored the other so that after 60 trials the evidence was equally strong (or equally weak) for both. People tended to favor the urn indicated by the first 30 draws, which is to say that the evidence in the first 30 draws was not counterbalanced by the evidence in the second 30 even though statistically it should have been. This is one illustration of what has been called a *primacy effect*. There is other evidence that how data influence one's evaluation of hypotheses can depend on when they are obtained in the hypothesis evaluation process (Chenzoff et al., 1960; H. C. A. Dale, 1968).

Studies have revealed a number of related effects. They all might be considered primacy effects in the sense that information that is processed early—or, perhaps more accurately, opinions that are formed from information that is processed early—seem to be given inappropriately great weight. These effects are discussed under such rubrics as persistence, commitment, and confirmation bias, and they may not all have the same basis.

The primacy effect is of considerable practical importance as it relates to judicial reasoning. Jurors are admonished *not* to form opinions as to the guilt or innocence of a defendant until all of the evidence has been presented and they have had a chance to deliberate about it. Complying with this admonition requires an ability to repress any tendency to form an initial opinion on the basis of partial information. Whether jurors can do this effectively is a matter of some doubt (G. P. Kramer, Kerr, & Carroll, 1990). Results of mock trial experiments indicate that mock jurors often come to favor a particular verdict early in the trial process (Devine & Ostrom, 1985; Weld & Danzig, 1940; Weld & Roff, 1938), and that they typically do reach at least a tentative verdict in their own minds before the deliberation process begins (Hastie, Penrod, & Pennington, 1983).

Additional Considerations

Demonstrating that a participant in an experiment has actually made use of Bayes's rule in solving a problem is complicated by the fact that one can sometimes get a non-Bayesian answer to a posterior probability question that would be close to the answer provided by application of Bayes's rule. Hoffrage (2000) points out, for example, that the ΔR rule (the hit rate minus the false-alarm rate) can sometimes yield answers close to those obtained from application of Bayes's rule. The ΔR rule has been found to be used sometimes by physicians and medical students and has been promoted as the correct way to estimate the covariation between two dichotomous variables (McKenzie, 1994).

Gigerenzer and Hoffrage (1995) argue, and present evidence, that people are more likely to reason in accordance with Bayes's rule if the pertinent information with which they have to deal is presented in frequency terms rather than as probabilities. People naturally think in terms of frequencies, they contend; the ability to think in probabilistic terms is a relatively recent and somewhat academic acquisition.

The question of whether Bayes's rule should be considered an appropriate norm against which to judge human reasoning, or more generally, that of what does constitute an appropriate norm, continues to be a matter of debate (L. J. Cohen, 1977, 1981; Gillies, 1973; Glymour, 1996; J. M. Harris, 1981; Macdonald, 1986). Non-Bayesian analyses generally can be applied to situations for which a Bayesian analysis is considered appropriate by proponents of a Bayesian approach. Gigerenzer et al. (1989), for example, have analyzed the taxicab problem (discussed in chap. 4) in terms of Neyman–Pearson statistics. Birnbaum (1983) has done so in terms of signal detection theory. As Gigerenzer et al. point out, concluding that people are irrational on the grounds that their behavior deviates from a particular normative model requires the assumption that that model is the only one that could constitute an appropriate norm. In the case of the taxicab problem and its various analogues, it seems that there are more models than one that could serve this function, so the most one can say of people's behavior is whether or not it is consistent with a particular view of what constitutes rational behavior in that situation.

But, given several alternative normative models, how is one to select among them? Should one be free to be a Bayesian on Mondays and Tuesdays, a Neyman–Pearson statistician on Wednesdays and Thursdays, and a Fisherian the remainder of the week? Each of the major schools of statistical reasoning has something to recommend it; and each has some very competent proponents who understand not only the theory they espouse, but the others as well. It may also be true, however, that many people who are irrevocably committed to a particular view are not familiar with the alternatives. Rationally deciding among alternative views requires some familiarity with all of them and of the assumptions on which they are based; one may well conclude that one view is more appropriate under some circumstances and another under others, but without a knowledge of the alternatives, it is not clear how a rational choice can be made. And in any case, ultimately one must decide what assumptions one is willing to make, and for this one has nowhere to go but one's intuitions.

THE VERDICT REGARDING STATISTICAL INTUITIONS

It is time to try to come to some conclusions. How good (or bad) are our intuitions regarding things probabilistic? According to several reviews of the experimental literature on failures in inductive reasoning, including Nisbett and L. Ross (1980), Einhorn and Hogarth (1981), Kahneman, Slovic, and Tversky (1982), Gilovich (1991), and Piatelli-Palmarini (1994), our reasoning under uncertainty or about statistical variables is subject to a variety of cognitive illusions and systematically goes wrong in numerous ways. Principles and relationships that are said often to be violated or ignored include the law of large numbers, the principle of regression to the mean, and the base-rate principle. To be sure, some investigators have presented evidence that people often do apply statistical principles effectively in dealing with problems for which statistical reasoning is appropriate (Jepson et al., 1983; Nisbett, Krantz, Jepson, & Fong, 1982; Nisbett et al., 1983), but , on balance, as Goldman (1986) notes, the work of the recent past on probability judgments appears to have been interpreted by most of those who have done or reviewed it as evidence of human irrationality.

With respect to the quality of probability judgments, G. N. Wright and Phillips (1980/1986) make a distinction among normative goodness, substantive goodness, and calibration. According to this distinction, normative goodness and substantive goodness reflect, respectively, conformity to the axioms of probability theory and consistency with what is known about the topic of judgment (Winkler & Murphy, 1968), and calibration refers to the degree to which judged events occur with a relative frequency that corresponds to their judged probability of occurrence (Lichtenstein et al., 1977). The adequacy of human reasoning has been challenged in all these respects.

Slovic et al. (1976) argue that, as intuitive statisticians, we have some serious deficiencies of a relatively fundamental sort: "The experimental results indicate that people systematically violate the principles of rational decision making when judging probabilities, making predictions, and otherwise attempting to cope with probabilistic tasks" (p. 169). They go on to say:

> Most of the discussions of "cognitive strain" and "limited capacity" that are derived from the study of problem solving and concept formation depict a person as a computer that has the right programs but cannot execute them properly because its central processor is too small. The biases from availability and anchoring certainly are congruent with this analogy. However, the misjudgment of sampling variability and the errors of prediction illustrate more serious deficiencies. Here we see that people's judgments of important probabilistic phenomena are not merely biased but are in violation of fundamental normative rules. Returning to the computer analogy, it appears that people lack the correct programs for many important judgmental tasks. (p.173)

Pennington and Hastie (1993) give the following litany of the ways in which human uncertainty assessment has been shown to be inconsistent with one or more of the traditional probability calculi:

> For example, the subjective probabilities of complementary hypotheses have been found not to sum to one (Edwards, 1962; Einhorn & Hogarth, 1985; [L. B.] Robinson & Hastie, 1985; Schum & [A. W.] Martin, 1980; van Wallendael, 1989; van Wallendael & Hastie, 1990); if certainty about one hypothesis increases, certainty about alternate hypotheses may remain constant; increase or decrease ([L. B.] Robinson & Hastie, 1985; Schum & [A. W.] Martin, 1980); hypotheses held with subjective certainty zero are frequently "revived" (Schum & [A. W.] Martin, 1980); the subjective certainty attached to a conjunction of events is frequently overestimated relative to the optimal combination of the component uncertainties (Bar-Hillel, 1973; Goldsmith, 1978); indeed, the subjective certainty attached to a conjunction of events may be assessed to be greater than the certainty of one or more of the component events (Leddo, Abelson, & Gross, 1984; Tversky & Kahneman, 1983); subjective certainty assessments may be too high under conditions where there is a high similarity between the pattern of evidence and a known standard, or when there is high internal consistency of the evidence even though the evidence is known or thought to be unreliable (Saks & Kidd, 1980; Schum, DuCharme, & DePitts, 1973; Schum & [A. W.] Martin, 1982; Tversky & Kahneman, 1974). (p. 213)

Nisbett and L. Ross (1980) contend that those studies that have supported the idea that people are rather good at intuitive statistics have tended to use "highly impoverished stimuli" (p. 74) and that biases are more likely to be seen when the stimuli that are used are "more interesting and more complex, when the boundaries of the relevant stimulus domains are less clear, and when recall, imagination, or inference provide the basis for the subjects' estimates" (p. 74).

Some assessments of our capabilities as probabilistic reasoners are damning indeed. Piattelli-Palmarini's (1989) is one case in point: "We know that our uneducated intuitions concerning even the simplest statistical phenomena are largely defective" (p. 9). S. J. Gould's (1993) is another: "Nothing is more unfamiliar or uncongenial to the human mind than thinking correctly about probabilities" (p. 280).

On the brighter side, several writers have argued either that human intuitions about probability are more sound than the more pessimistic assessments of experimental results suggest or that many of the apparent failures of probabilistic reasoning in the laboratory may not be indicative of how people perform when they face less contrived real-world situations that require probabilistic reasoning. Some have pointed out characteristics of many laboratory experiments that limit the generalizability of some of the conclusions that have been drawn from their results to real-world situations of interest. Ayton et al. (1991a), for example, argue that "a large number of experiments purportedly showing that humans are illogical or poor intuitive statisticians may be examining performance on inappropriate tasks.... It remains possible that, within the usual naturally occurring framework for human induction, performance is highly successful" (p. 227).

Nisbett et al. (1983) argue that people do possess and use inferential intuitions that resemble formal statistical procedures, and that they do so effectively at least under certain conditions. They present evidence for the hypothesis that whether people use appropriate statistical reasoning depends on such factors as the clarity of the sample space and the sampling process, whether the reasoner recognizes the role of chance in the situation of interest, and whether there is a cultural or subcultural prescription to reason statistically about the events in question.

Holland, Holyoak, Nisbett, and Thagard (1986) suggest that people are likely to reason more statistically—to generalize less strongly from extreme events, to be less inclined toward causal explanations when outcomes are different on superficially similar occasions—when reasoning in domains in which variability and randomness are relatively easy to assess and to reason less statistically in domains in which such assessments are more difficult to make. This is one of the ways in which the inferential thinking one does about a domain is affected by one's factual knowledge of the domain.

Jepson et al. (1983) found that people are more likely to give statistical answers for problems the probablistic nature of which is relatively obvious (e.g., problems dealing with lotteries) than for problems whose probabilistic nature is less apparent. Such findings have led Nisbett and his colleagues to conclude that people possess intuitive and abstract versions of statistical rules, which they refer to as "statistical heuristics." Failure sometimes to use such rules is seen as resulting from failure to encode problems in such a way as to evoke the rules or as a consequence of the evocation of competing heuristics (Fong, Krantz, & Nisbett, 1986; Nisbett et al., 1983).

Perhaps the most serious criticism that has been made of the work on intuitive statistics and, in particular, of the numerous studies that have been used to support the general conclusion that people are poorly calibrated and tend to be overconfident of their judgments, is that the great majority of these studies have been done with college students as participants who have been asked to make judgments on matters about which they are not highly knowledgable or for which they have little or no intrinsic interest. The applicability of findings from these studies to people making judgments in their areas of expertise has been challenged (J. J. Christensen-Szalanski & Beach, 1984; Hogarth, 1975; Pitz, 1974). Apparently, when people make many motivated judgments of a similar type in situations that provide them with feedback as to the accuracy of those judgments (e.g., weather forecasting or contract bridge), they can become very well calibrated (Keren, 1987; Murphy & Winkler, 1977; Wallsten & Budescu, 1980). On the other hand, there is evidence that professionals can be poorly calibrated even when making judgments in their fields (Dawes, 1988; L. R. Goldberg, 1959). It appears that overconfidence is likely to be an occupational hazard in fields that are not sufficiently objective or well developed to provide practitioners with unambiguous feedback regarding the accuracy of specific judgments within them.

Gigerenzer (1991b) challenges the widely held idea that people are not generally competent intuitive statisticians and questions the validity of much of the experimental evidence from which this idea has received support. In particular, he objects to the practice of treating one prescription for reasoning under uncertainty as though it were the only legitimate way to proceed:

> Good judgment under uncertainty is more than mechanically applying a formula, such as Bayes's theorem, to a real-world problem. The intuitive statistician, like his professional counterpart, must first check the structure of the environment (or of a problem) in order to decide whether to apply a statistical algorithm at all, and if so, which.... There is no good (applied) probabilistic reasoning that ignores the structure of the environment and mechanically uses only *one* (usually mathematically convenient) algorithm. (p. 106)

With respect to Bayes's theorem in particular, Gigerenzer (1991b) points out that its successful application depends on several structural assumptions:

"independence of successive drawings, random sampling, an exhaustive and mutually exclusive set of hypotheses, and independence between prior probabilities and likelihoods" (p. 107). If one is to apply Bayes's theorem appropriately to real-world problems, one must make judgments regarding whether situations of interest have the assumed characteristics, and such judgments must draw on one's knowledge of the world and on inferences or conjectures that can be made therefrom.

In short, the literature on probabilistic reasoning is conflicted. On the one hand are the numerous studies that ostensibly have shown that people often reason poorly when thinking about probabilistic or statistical situations, and that even people who are knowledgeable with respect to statistics often do not apply that knowledge appropriately to practical problems. On the other hand are those studies that show that many people trained in statistics do make appropriate use of their knowledge and that even people without such training often apply relatively sound intuitions to probabilistic and statistical problems, and claims that many of the results that have been taken as evidence of faulty probabilistic intuitions could be interpreted in ways that are less denigrative of human rationality.

It should not escape our notice that the claims that people often give evidence of poor statistical intuitions and that they often give evidence of sound statistical intuitions are not mutually exclusive; they both could be correct. Moreover, it is not impossible that some people who have been schooled in probability theory and statistics typically apply appropriately what they have learned and that others do not, or that the same individuals, learned with respect to these disciplines, sometimes apply what they have learned effectively and sometimes fail to do so.

Irrationality or Lack of Mathematical Knowledge?

Goldman (1986) questions whether the results of experiments on probability judgments really constitute evidence of irrationality. He notes the controversial nature of the interpretation of probability statements and uncertainty as to what should be considered normative principles of probability judgment. Though acknowledging the value of probability theory and statistics as intellectual tools and methods, he questions whether competence with these tools should be a requirement for rationality: "Probability theory is a branch of mathematics, like other branches of mathematics. Failure to have learned or mastered other branches of mathematics is not normally taken as a shortcoming in rationality. Nor is the failure to recognize every concrete application of such branches. By parity of treatment, it would be wrong to view every deficiency in grasping or applying probability theory as a specimen of irrationality" (p. 316).

To press the point, Goldman (1986) notes that the modern concept of probability began to be clarified only in the 17th century, and was not axiomatized until the 20th century. It is hardly surprising, he argues, if untrained people should find it difficult to arrive at an adequate conception of probability by themselves. Goldman is careful to note that he is not arguing that people are natively rational in matters of probability, only that the question cannot now be answered and, in particular, that experimental results showing that people often violate the prescriptions of probability theory in their reasoning are not compelling evidence of irrationality.

On the question of what would constitute irrationality in this context, Goldman (1986) suggests that it would have to be a defect in basic processes as distinct from failure to have mastered certain intellectual tools, and that the defective processes would have to be reasoning processes as distinct from, say, memory retrieval processes. Regarding in particular the processes that yield "erroneous" probability judgments of the type demonstrated by numerous investigators over the past few years, whether these processes should be considered irrational depends on how reliably they produce correct judgments considering all the situations in which they are typically applied. It is at least conceivable that these processes are usually reliable in situations less contrived than those often used in experiments.

There is also a question of what it means to understand probability and the principles underlying statistics. Paulos (1998) makes a distinction between formal facility with these concepts and an understanding of them at an intuitive level. One may become very adept at games of chance—poker, bridge—giving evidence of having acquired the ability to make choices that are prescribed by the probabilities of the various possible outcomes, without being able to verbalize the principles on which one is operating in mathematical terms.

The Question of Criterion

> In most studies that claim to have demonstrated human errors, biases, and shortcomings, no argument is given to explain why the statistical rule *is* rational, nor is rationality independently defined. (King, 1992, p. 179)

Any claim of irrationality presumes some standard with which the behavior that is considered to be irrational is compared and found wanting. In the case of statistical or probabilistic thinking, a standard that has generally been applied in the Western world, although with some waxing and waning of popularity over time (Martignon & Laskey, 1999), is probability theory, and in particular, probability theory as developed from the formulations of such 17th-century thinkers as Pascal and Fermat; judgments or decisions are said to be irrational when they are inconsistent with what "Pascalian" probability theory prescribes.

A forceful argument against the conclusion that people generally exhibit irrationality in the sense of being at odds with probability theory has been made by Gigerenzer (1991b, 1994). The argument rests on the claim that there is no single normative theory of probability that is recognized as such by all mathematicians who work in this area. There are rather several views of what probability means and of what the basic tenets of a normative theory should be, so the worst that can be said about any particular judgment is that it is irrational according to this or that theory of probability, but what is irrational according to *this* theory may be exactly right according to *that* one.

Gigerenzer (1991b) argues that much of the psychological literature on cognitive illusions and biases ignores this fact, that it assumes the existence of an unequivocal normative theory, and that the theory it treats as definitive is not considered as such by all major probability theorists:

> What is called in the heuristics and biases literature the "normative theory of probability" or the like is in fact a very narrow kind of neo-Bayesian view that is shared by some theoretical economists and cognitive psychologists, and to a lesser degree by practitioners in business, law, and artificial intelligence. It is *not* shared by proponents of the frequentist view of probability that dominates today's statistics departments, nor by proponents of many other views; it is not even shared by all Bayesians.... By this narrow standard of "correct" probabilistic reasoning, the most distinguished probabilists and statisticians of our century—figures of the stature of Richard von Mises and Jerzy Neyman—would be guilty of "biases" in probabilistic reasoning. (p. 87)

More generally, Gigerenzer et al. (1989) have been critical of the elevation by psychologists of probability theory "as a mathematical codification of rational belief and action in uncertain situations" (p. 226). They point out that, unlike the 18th-century probabilists who had been willing to revise their mathematics to fit better the dictates of common reason when they ran into the St. Petersburg problem, "the twentieth-century psychologists had come so to revere the mathematical theory of probability and statistics that they instead insisted that common reason be reformed to fit the mathematics" (p. 226).

The objective of research on probabilistic reasoning should not be to attempt to explain the difference between people's judgments and "normative" performance as represented by a particular Bayesian prescription, Gigerenzer (1991b) argues, but rather to explain people's judgments. More progress will be made, he suggests, by viewing competing statistical theories as the bases of competing explanatory models than by pretending that statistics speaks with one voice:

> The history of probability theory, with all its changes in the interpretation of probability, in the meaning of "descriptive" and "normative," ... should warn us to be cautious in using one formal method as *the* norm, against which such judgments are denigrated as irrational, independent of their content and context. A rational mind may be more than the kind of intuitive statistician who mechani-

> cally applies the same formula (be it expected utility, Bayes' theorem, or some other) to all contents and in all contexts. But such a rational mind is much harder to define. (Gigerenzer et al., 1989, p. 232)

A similar position is taken by L. J. Cohen (1981), who not only distinguishes among several normative theories of probability, not all of which derive from the seminal work of Pascal and Fermat, but argues that the establishment of a normative theory does not prove that theory to be appropriate for judging the probabilistic reasoning of people in everyday life:

> It is one thing to establish one or more probabilistic interpretations for the calculus of chance, and quite another to show that the resultant theory applies to some or all of the probability judgments that are made in everyday reasoning. In order to discover what criteria of probability are appropriate for the evaluation of lay reasoning we have to investigate what judgments of probability are intuitively acceptable to lay adults and what rational constraints these judgments are supposed to place on one another. (p. 319)

L. J. Cohen (1982) contrasts the "Preconceived Norm Method" with the "Norm Extraction Method" as approaches to the interpretation of experimental data about probability judgments. An investigator using the preconceived norm method "tacitly assumes that the problem-task set to his or her subjects is correctly soluble only in terms of some academically well-regarded conception of probabilities that he or she has in mind. The investigator therefore evaluates the subjects' performance for correctness or incorrectness by a technique of assessment or estimation that is appropriate to this mode of conception" (p. 251). Underlying the norm extraction method is the assumption that "unless their judgment is clouded at the time by wishful thinking, forgetfulness, inattentiveness, low intelligence, immaturity, senility, or some other competence-inhibiting factor, all subjects reason correctly about probability: none are programmed to commit fallacies or indulge in illusions" (p. 251). According to this view, the purpose of experimentation is not to find out how people's reasoning deviates from the prescriptions of preconceived norms but to discover the conceptions of probability they apply to problems of specific types:

> In short, the Preconceived Norm Method assumes a standard conception of probability, imputes its acceptance to the subjects, and hypothesizes either faulty programming or temporary causes of malfunction in order to account for estimates that are erroneous in terms of that conception: the Norm Extraction Method hypothesizes about the subjects' conception of probability and their mode of assessing it, on the assumption that unless affected by temporary or adventitious causes of error their judgments are correct. (p. 252)

This is not to say that people cannot contradict themselves or make invalid deductions. But when they do these things, the norms they violate, Cohen argues, can be discovered from their own intuitions.

L. J. Cohen (1979) rejects, in particular, the idea that Pascalian probability theory is the only legitimate framework within which to reason about uncertain predictions. He argues that one normative alternative to the Pascalian formulation that is widely used in science can be traced to the writings of Francis Bacon. Cohen argues further that before one can legitimately classify an instance of reasoning about probability as fallacious, one must know the theoretical framework within which the reasoning is being done. What would be considered fallacious from a Pascalian perspective would not necessarily be seen as fallacious from a Baconian point of view. Many of the alleged demonstrations of human fallibility in probabilistic reasoning are unconvincing, Cohen argues, because they *assume* a particular perspective—the Pascalian one—that people may not always use. Pascalian and Baconian judgments may differ from one another in any particular circumstance, which is not to say that at least one must be wrong. Neither, he contends, implies the falsity of the other.

The idea that two different judgments of probability regarding the same situation could both be right may be difficult for the reader to accept. I find it hard to accept. Could we not settle the issue in favor of one or the other judgment—or possibly against both of them—by doing an appropriate experiment? If the probability of a specified outcome of an uncertain event is P_P and P_B, according to the Pascalian and Baconian views, respectively, let us simply observe the event a large number of times and determine the proportion of instances on which the specified outcome occurs. But, at best, this works only for events of the sort that can be observed a large number of times, and many of the events about which we wish to reason probabilistically are not this type. Even when it is possible, both in theory and in practice, to observe an event many times, what it means to say that the probability of a specified outcome on *a specific occurrence of that event* is thus and so is a debatable question. When I say that I believe that the probability that the toss of a fair die will produce a three with probability 1/6, I am sure I mean that I believe that three will come up on approximately 1/6 of a large number of tosses. But suppose that the die is to be tossed just once. What I mean by saying that I believe the probability that it will come up three on that toss is 1/6 is not quite so clear.

L. J. Cohen (1970, 1977) has developed the Baconian view of probability in some detail. According to this view, probability has to do primarily with judgments of the "inductive reliability of generalizations," and is focused on the inductive establishment of causal laws. Cohen argues that the experimental method of modern science is essentially Baconian in that it seeks to establish the inductive reliability of generalizations, which we commonly refer to as scientific laws. The Anglo-American legal system also endorses Baconian reasoning, in Cohen's view, by virtue of its commitment to adjudication on the basis of the totality of the case-specific evidence that is before the court. No court, he suggests, would convict a person in a civil suit solely on the grounds that he belongs to a group over half of the members of which are known to have committed the crime of which he is accused (base-rate information). He points out too that if, in criminal cases, the injunction on jurors to assume a defendant is innocent until proven guilty is interpreted as the assignment of zero as the prior probability of guilt, then a guilty verdict within a Pascalian–Bayesian framework is impossible, because in this system prior probabilities of zero inevitably produce posterior probabilities of zero, independently of what the conditional probabilities are. None of this is to argue that Pascalian probabilities cannot enter into forensic reasoning, but only that cases can seldom be decided on the basis of them alone.

Many of the experimental findings that have been interpreted as evidence of fallacious reasoning—application of the representativeness heuristic, for example—may be seen as fallacious, Cohen argues, if one assumes that reasoners were working within the framework of Pascalian probability but not if they were approaching the problems from a Baconian perspective. As to which perspective reasoners *should* use in specific instances, Cohen argues that the charitable thing for experimenters to do is to assume that participants in their experiments are at least as rational as themselves and to interpret their performance on probabilistic reasoning tasks in whichever ways do not require the assumption that they are reasoning fallaciously: "One hypothesis which seems to fit the available evidence is that people inexpert in statistical theory tend to apply Baconian patterns of reasoning instead, and to apply these correctly whenever they have an opportunity to make the probability in question depend on the amount of inductively relevant evidence that is offered" (L. J. Cohen, 1979, p. 397).

In short, Cohen argues that many uncertain situations can be assessed from either a Pascalian or a Baconian point of view, that neither perspective is always clearly more appropriate than the other, and that before one can legitimately classify an instance of probabilistic reasoning as fallacious, one must know which perspective the reasoner is taking. Cohen allows that there are situations in which a particular view is correct and the alternative is not—the Pascalian view is the only appropriate one, for example, when the statement of a problem makes it clear that the probabilities involved are equivalent to long-term percentages or relative frequencies—and he argues that the evidence that people characteristically reason fallaciously in such cases is not compelling, which is not to deny that certain types of fallacies may be fairly common.

One of the more common criticisms of research focused on the identification of common biases in probabilistic thinking, or on ways in which such thinking commonly violates normative principles, is that the situations investigated sometimes are contrived so as to ensure that the application of heuristics or principles that are effective in everyday life will be inappropriate in the experimental context (Ayton et al., 1991a, 1991b; Goldman, 1986; Winkler & Murphy, 1973). Moreover, what may appear to be an irrational bias in one context may be functional in another. For example, the negative-recency bias (see chap. 2) would be a useful bias to have in any context in which it is important to be able to detect nonrandomness in sequences and the most likely form that departures from randomness take is that of excessive repetitions or an excess of repetitions is more important to detect than an excess of alternations (Lopes, 1982). Navon (1978) notes ways in which the behavior that is typically considered evidence of excessive conservatism in Bayesian decision experiments can be beneficial in real-world contexts.

Uncertainties About Task Perception

In psychological experiments it is important to distinguish two tasks: the one the experimenter intends for the participants to perform, and the one the participants actually perform. One hopes the two are identical, or at least very similar, but there is room for doubt in many instances. Unfortunately, it is usually simply assumed that participants perceive the task as does the experimenter and the results are interpreted without consideration of the possibility that this is not the case. It may be that experimenters are sometimes unaware of the possibility that a task could be perceived differently from the way they perceive it, but if it can be and even a small percentage of the participants do perceive it in an alternate way, this could invalidate conclusions drawn from the results.

The point is illustrated by some experiments on the perception of randomness. Failure to specify whether sampling is to be with replacement or without leaves open the possibility of participants operating on either assumption, and to the extent that some of them make the assumption that is inconsistent with that made by the experimenter, the results are likely to be misinterpreted. The importance of this point is demonstrated by an experiment of Winefield (1966) in which he showed that the usual negative-recency effect was not obtained in a card-guessing task when participants witnessed the reinsertion of each drawn card into the deck before the deck was shuffled and the next card drawn. The sensitivity of people's performance of probabilistic reasoning tasks to differences in instructions, including sometimes fairly subtle ones, should give pause in the drawing of conclusions about human reasoning foibles on the basis of performance of tasks when there is any question of how participants might have perceived those tasks.

There is also the problem that different participants can sometimes interpret the same instructions differently. There is some evidence, for example, that when asked to estimate the probability of a cause given an effect in experiments on diagnostic judgment—say the probability of a rare disease given a specific symptom—they have actually estimated the probability of the effect given the cause (J. R. Anderson, 1990). According to Eddy (1982), even physicians commonly confuse the probability of a disease state given a positive diagnostic test result with the probability of getting a positive test result given that the patient has the disease for which the test is intended to be diagnostic. This shows a lack of statistical sophistication, but it is not clear whether it reflects only a terminological confusion or a deeper reasoning problem.

Some of the errors that people make in estimating correlations, contingencies, and other statistical relationships also may be due to a misunderstanding of the problem on which they are working. Problems can be stated in more or less straightforward or obscure ways and some of the situations that have been used in experimental studies have been contrived for the purpose of making it especially easy for people to blunder in one way or another. But even when instructions are intended to be straightforward and clear, they sometimes can be ambiguous or incomplete. This is seen by some researchers to have been especially problematic in studies of probabilistic reasoning (Bar-Hillel & Falk, 1982; Falk, 1992; Margolis, 1987; Nickerson, 1996).

Difficulties stemming from the ambiguity of problem statements are compounded by the fact that ambiguities are typically not recognized as such. Social psychologists speak of an *assimilation effect* in reference to the finding that when people encounter ambiguous information they interpret it in terms of the concept that is most accessible at the time (Schwarz, 1995). This effect has often been observed in social contexts in which people have been primed to see ambiguous behavior in either one of two ways (e.g., persistent or stubborn, friendly or hostile) (Wyer & Srull, 1989). The typical finding is that people interpret the behavior in the primed way and fail to see the alternative interpretation as a possibility.

Bases of Erroneous Statistical or Probabilistic Thinking

The ambiguity of many of the situations used to study probabilistic reasoning in the laboratory notwithstanding, the evidence seems sufficient to warrant the conclusion that many people, including people with post–high school education, either lack the ability to do even quite elementary statistical reasoning, or are disinclined to make the effort necessary to apply what they know about probability and statistics to the problem of reasoning under uncertainty.

In some cases, inability to deal effectively with problems of probability and statistics appears to stem from a lack of understanding of some of the basic principles that define these disciplines. The following examples illustrate the point. Pollatsek, Konold, Well, and Lima (1984) had college students estimate (a) the mean of a random sample of 10 scores consisting of 9 unknown scores and 1 known score that differed substantially from the population mean and (b) the mean of the 9 unknown scores. About 40% of the subjects gave the population mean as the answer in both cases. When undergraduates were asked to judge whether the rate of teenage pregnancies was changing at a hypothetical high school, they tended to base their judgments solely on changes in the number of pregnancies and to ignore changes in the school population, unless population size was made highly salient in the statement of the problem (Silka & Albright, 1983).

Pollard and J. T. Richardson (1987) suggest that difficulties with conditional probabilities could stem in part from the assumption that a syllogistic form that is valid with deterministic statements is necessarily valid also when used with probabilistic statements. They illustrate how this assumption could get one into trouble with the following examples:

> If H then not D
>
> *D*
>
> Therefore, not H

and

> If H then not very likely D
>
> *D*
>
> Therefore, not very likely H.

The first of these forms is a valid *modus tollens* argument. The second form looks fine, but that it is not valid is easily seen by the following instantiation of it:

> If this person is an American, this person is a not very likely to be a member of Congress
>
> *This person is a member of Congress*
>
> Therefore, this person is not very likely to be an American.

This confusion is often seen in misinterpretations of the *p* values obtained in the results of null-hypothesis significance testing (Nickerson, 2000). According to the theory behind null-hypothesis testing, if the null hypothesis is true, one is unlikely to obtain a *p* value as small as .05, but the obtaining of a *p* as small as .05 is often taken as evidence that the null hypothesis is unlikely to be true:

If H_0 is true then one is not very likely to get $p < .05$

p < *.05*

Therefore, H_0 is not very likely to be true

Errors to which the gambler's fallacy would be expected to give rise appear to be based, in some cases at least, on misconceptions about what a random sequence or a noncontingent relationship is (C. R. Peterson, 1980). Peterson argues that participants in psychological experiments sometimes fail to perceive noncontingent events as noncontingent because the expectations they bring to the experiment preclude the assumption that what is occurring is random. When one does not make the assumption that the process that is determining the outcome of an uncertain event is completely random, then it is questionable whether what is usually deemed an instance of the gambler's fallacy is appropriately considered a fallacy (L. J. Cohen, 1982); moreover, in many experimental situations what are represented to participants as random processes are really not random.

Gilovich (1991) also argues that some of our reasoning difficulties stem from the strong tendency we seem to have of perceiving even random events and patterns as having some degree of structure: "Our difficulty in accurately recognizing random arrangements of events can lead us to believe things that are not true—to believe something is systematic, ordered, and 'real' when it is really random, chaotic, and illusory" (p. 21). This tendency is exacerbated by the fact that once we believe a phenomenon exists we are very good at inventing reasons for why it should exist.

People untutored in probability theory may fail to understand that the probabilities of a set of mutually exclusive possibilities cannot total more than 1 or that the elimination of some of the alternatives from among such a set has implications for the probabilities associated with the remaining alternatives. When told, in one study of statistical reasoning, that a certain one of a few suspects in a murder mystery was not the culprit, participants tended to fail to revise the probabilities assigned to the remaining suspects (L. B. Robinson & Hastie, 1985).

Sometimes it is difficult to tell whether inappropriate decisions based on the probabilities of outcomes from random sampling are better attributed to faulty intuitions regarding fundamental aspects of probability theory or to an inability to compute or estimate what the probabilities of specific outcomes are. This comment applies, for example, to the finding of L. J. Cohen and Chesnick

(1970) that some people prefer a lottery in which they draw 1 ticket from a population of 10 tickets to the opportunity to have as many as 20 draws (with replacement) from a population of 100 tickets.

Evans (1989) argues that what appears to be faulty reasoning under uncertainty results in part from faulty intuitions, but more from a generally passive approach to reasoning, whereby people tend to consider only what is explicitly brought to their attention. Even correct intuitions, he suggests, are readily overpowered by irrelevancies when the latter are made salient by the way in which problems are posed to them. This view is compatible with Baron's (1985, 1988) attribution of faulty reasoning to insufficient search. In both cases, lack of effort—as distinct from lack of knowledge—is seen as a major cause of poor reasoning. That performance on many tasks used in psychological experiments may elicit less-than-maximum effort by the participants is a highly plausible possibility. That students who participate to satisfy a course requirement, or to obtain extra credit, or even for cash payment will invariably make a best effort on tasks in which they may have little or no intrinsic interest is an assumption that is perhaps too easily made uncritically. This is not to suggest that all experimental tasks fit this description for all participants, but certainly many do, and that suffices to make one cautious about generalizing the results of many experiments to situations in which people are highly motivated to do their best on tasks because they have a meaningful stake in the consequences of their performance.

Although it is perhaps somewhat ironic to look to psychological experiments for evidence of this assertion, a number of studies have shown that the amount of effort participants devote to the analysis and interpretation of information presented to them depends, at least in part, on the degree of their personal involvement with the issues to which the information pertains (Chaiken, 1980; Harkness, DeBono, & Borgida, 1985). Highly involved people appear to process messages and arguments at a deeper level and more extensively than less involved people, who are likely to be influenced by relatively superficial or surface features (Borgida & Howard-Pitney, 1983; Petty & Cacioppo, 1979; Petty, Cacioppo, & Goldman, 1981; Showers & Cantor, 1985).

As Harkness et al. (1985) point out, assuming some modest correlation between complexity of information-processing strategies and the quality of the outcomes of those strategies in natural task environments, it makes sense for people to use more complex strategies when they are more involved—when they have more at stake. These investigators demonstrated a relationship between involvement and complexity of strategies used in a study in which data about a potential dating partner were processed by participants, some of whom were told they would actually be dating the individual several times and some of whom were not given this expectation. Although the highly involved participants in this study used more complex strategies than did the less involved participants to analyze contingency data, the most common approach among them was not the normatively optimal one. Harkness et al. note that the tendency for uninvolved participants to use easy and superficial strategies can be seen as an indication of rationality—as evidence of mindfulness in deciding when to be mindless: "Suboptimal performance on a trivial task may be a marker of intellect" (p. 32).

In short, investigators have identified many possible bases of what appears to be faulty reasoning under uncertainty. I say what "appears to be faulty reasoning" because some of the bases that have been suggested might better be considered misperceptions, limitations of knowledge, or failures of motivation than of faulty reasoning per se. Causes that have been suggested include confusion about what one has been asked to do in an experimental situation, insensitivity to such basic distinctions as that between frequency and relative frequency, lack of understanding of certain critical concepts such as those of randomness and contingency, failure to understand that the probabilities of an exhaustive set of mutually exclusive possibilities must sum to 1, computational limitations, misapplication of heuristic rules, and a tendency to approach problems too passively and hence to fail to make the effort necessary to exploit one's reasoning capabilities fully.

The Battle of Intuitions

Much of the research that has been reviewed in this book has focused on ways in which probabilistic reasoning appears to go astray. But the judgment that any particular instance of reasoning is faulty presupposes one has a standard against which to make that judgment. As Larkey et al. (1989) put it, "Attributing error in reasoning about chance processes requires at the outset that you know the correct model for the observations about which subjects are reasoning. Before you can identify errors in reasoning and explain those errors as the product of a particular style of erroneous reasoning, you must first know the correct reasoning. It is much easier to know the correct model in an experimental setting than in a natural setting" (p. 30).

A general conclusion that has been drawn from the work that has been considered here is that people's intuitions about probability and related concepts—chance, randomness, covariation—are faulty. However, what to make of this conclusion is called into question by the fact that highly knowledgeable people disagree, rather strongly on occasion, as to what should be considered fallacious and faulty and what should not (L. J. Cohen, 1979, 1980; Gigerenzer, 1994; Kahneman & Tversky, 1979a).

What is clear is that there is no escaping appeal to intuition. When the intuitions that underlie the probabilistic reasoning of participants in experiments are judged to be faulty by the experimenters, those judgments are themselves based on other intuitions about probability—the experimenters' intuitions, informed by one or another school of thought that represents the intuitions of other scholars. And not all writers have the same intuitions in this regard.

We should not forget the essential role that intuition played in the development of probability theory in the first place. As Daston (1987b) puts it: "Because its practitioners understood probability theory as a mathematical codification of good sense, the right answers to these questions were those that seconded the intuitions of reasonable men" (p. 298). And again, "When mathematical results clashed with the practice of reasonable men, the eighteenth-century probabilists consistently rearranged or modified the mathematics to reconcile the two" (p. 298). Or, as Gigerenzer et al. (1989) put it, probability theory was meant by its developers to capture, not to correct, reasonable intuitions.

In chapter 1 of this book, we noted the way Pascal and Fermat sharpened their respective intuitions in attempting to converge on a solution to the problem on which they were working that they would both consider correct. Daston (1987b) refers to the St. Petersburg paradox to illustrate how intuitions could be modified as a consequence of grappling with problems: "If the conventional solution of the St. Petersburg problem ran counter to the judgment of reasonable men, probabilists reexamined their definitions of expectation. In the minds of eighteenth-century mathematicians, there did not exist any theory of probabilities disembodied of subject matter" (p. 298).

So where does this leave us? To what do we as individuals appeal when we try to judge the merits of the differing—and mutually incompatible—points of view defended by the experts? We have no place to go but to our own intuitions. We have no choice but to accept those arguments that we find most intuitively compelling. But this is as it should be. If we are rational creatures, we, each of us individually, and not the experts, are responsible for our individual beliefs.

This does not mean that all possible views are equally correct, or that dialogue and debate are pointless. What one finds intuitively compelling can change as a consequence of the acquisition of information and exposure to new views and arguments in their support. But it does mean that the only legitimate defense that I can offer for holding a particular view at any given time is that *I* find that view to be the most consistent with the evidence at my disposal as judged by my own intuitions. If I adopt a belief about something of which I have little direct knowledge because it represents what certain experts believe, the decision to do so is my responsibility, and I must appeal to my intuitions—hopefully informed by some evidence of the competence and integrity of the experts in question—in making it.

EFFICACY OF EDUCATION AND TRAINING

Some of the studies reviewed here have been taken as evidence that even people who have had considerable exposure to probability theory and statistics, sometimes at postgraduate levels, reason poorly, on occasion, about random variables and processes. It does not follow, of course, that education and training are to no avail with respect to improving people's ability to reason under uncertainty; the fact that people who have had instruction in a particular discipline do not always behave in ways that are consistent with what they were taught is not compelling evidence that the teaching of that discipline is futile.

Several investigators have noted differences in the way people from different cultures assess probabilities and think about uncertain events (Phillips & G. N. Wright, 1977; G. N. Wright & Phillips, 1980/1986; Wright et al., 1978). These differences are not well understood. They do constitute evidence, however, that if there are any universal culture-free intuitions about probabilistic processes, there are also aspects of our conceptions of probability and its manifestation in the world that are learned, either by covert assimilation of viewpoints that are prevalent in one's culture or as a consequence of overt efforts to teach and inform.

What is the evidence with respect to the efficacy of instruction in probability and statistics relative to the quality of reasoning about matters probabilistic and statistical? What is known about the effectiveness of various approaches to the teaching of probabilistic and statistical reasoning? Has research shown that there are better techniques to improve reasoning under uncertainty than those being used?

One evidence of the effectiveness of instruction is the fact that even unexceptional students today are able to solve probabilistic and statistical problems that baffled accomplished mathematicians of previous centuries, thanks to the availability of a codified body of theory that can be taught. It is easy for us for whom this theory is available to overlook the fact that many problems that are quite simple from today's perspective were very difficult for people who did not have the benefit of the theory. Even the intellectual giants sometimes had difficulties with what today would be seen as relatively straightforward problems. Leibniz, for example, in his *Dissertatio de Arte Combinatoria,* which was published in 1666, gave the probability of tossing a 12 with two dice as equal to the probability of tossing an 11, the argument being that both numbers can be partitioned into two numbers between 1 and 6 in one and only one way (12 can be partitioned into 6 and 6, and 11 into 5 and 6) (Todhunter, 1865/2001, p. 48). What Leibniz overlooked, of course, was the fact that with a pair of dice, 11 can be obtained with a 5 and 6 in two ways (5 on the first die and 6 on the second, or vice versa) whereas 6 and 6 can be obtained in only one. We would not expect this kind of error to be committed today by anyone who had studied, and understood, basic combinatorics.

In his extensive review of work on probability theory from the time of Pascal to that of Laplace, Todhunter (1865/2001) gives many examples of the intuitions—and calculations—of eminent mathematicians that were later shown to be wrong. Perhaps the most striking example of a distinguished mathematician several of whose ideas about probability have attracted attention because of their untenability is D'Alembert. Todhunter says of him, "This great mathematician is known in the history of the Theory of Probability for his opposition to the opinions generally received; his high reputation in science, philosophy, and literature have secured an amount of attention for his paradoxes and errors which they would not have gained if they had proceeded from a less distinguished writer" (p. 258). D'Alembert believed, for example, that the larger the number of consecutive tosses of heads with a coin, the larger the probability of tails on the next throw (he was not the only mathematician of note to believe this); that the probability of getting a head at least once in two tosses of a coin was 2/3 (although the strength of this belief appears to have diminished somewhat over time); that the outcome probabilities associated with tossing a coin n times are different from those associated with the tossing of each of *n* coins once.

Todhunter (1865/2001), who was unsparing in his exposure of what he deemed suboptimal reasoning, incorrect computation, and obscure exposition by the major contributors to probability theory through the early 19th century, saved his most scathing assessment for Condorcet:

> We must state at once that Condorcet's work is excessively difficult; the difficulty does not lie in the mathematical investigations, but in the expressions which are employed to introduce these investigations and to state their results: it is in many cases almost impossible to discover what Condorcet means to say. The obscurity and self contradiction are without any parallel, so far as our experience of mathematical works extends; some examples will be given in the course of our analysis, but no amount of examples can convey an adequate impression of the extent of the evils. We believe that the work has been very little studied, for we have not observed any recognition of the repulsive peculiarities by which it is so undesirably distinguished. (p. 352)

Despite this assessment, Todhunter devotes 60 pages to a discussion, often critical, of Condorcet's work, which is approximately the same as he devotes to the work of both Montmort and De Moivre and about 10 pages more than he gives to that of Laplace. His parting shot regarding Condorcet: "Condorcet seems really to have fancied that valuable results could be obtained from any data, however imperfect, by using formulae with an adequate supply of signs of integration" (p. 410).

To the extent that inappropriate reasoning under uncertainty is due to a lack of understanding of certain basic probabilistic and statistical concepts, this should be correctable by instruction. Even a small amount of training has been shown to be effective, for example, in reducing the incidence or degree of illusory correlation (L. J. Chapman, 1967; Waller & Keeley, 1978). Nisbett and his colleagues have presented evidence that formal training in statistics has a positive effect on the quality of reasoning that people do on probabilistic and statistical problems (Nisbett, Fong, Lehman, & Cheng, 1987; Nisbett et al., 1983).

The teaching of probability and statistics should, I believe, have a high priority throughout the educational process, and perhaps especially during relatively early years. Throughout life it is necessary to reason about uncertain events, to make decisions when the information one has about the consequences of specific choices among one's alternatives is incomplete and probabilistic. A prerequisite of rational behavior in such situations is some understanding of probability and statistics and an ability to think in these terms. Inability to think probabilistically necessitates thinking in polar, dichotomous terms and maintaining oversimplified views of natural and social phenomena. Misconceptions about probabilistic events that appear to be quite pervasive and difficult to correct among adults might be prevented if students were taught to think in probabilistic terms while their thinking patterns are still in the formative and more fluid stages of development. I am aware of no evidence that probabilistic thinking is beyond the grasp of grade school students.

I have already noted that Nisbett et al. (1983) have presented evidence for the hypothesis that whether people use appropriate statistical reasoning depends on such factors as the clarity of the sample space and the sampling process, whether the reasoner recognizes the role of chance in the situation of interest, and whether there is a cultural or subcultural prescription to reason statistically about the events in question. Cultural and subcultural prescriptions presumably derive largely from educational institutions and practices. Probabilistic or statistical reasoning is unlikely to be pervasive in a society unless it gets the emphasis it deserves in the society's educational system.

SUMMARY

People are able to estimate statistical properties of events moderately well, but the accuracy of these estimates can be influenced by irrelevancies and biases of various sorts.

People have intuitions about the probabilities of events. They recognize some things, say a family of six children being all girls, to be less probable than others, say a family of four girls and two boys. These intuitions work fairly well in many everyday contexts; they are not very precise, however, and people generally would not be able to provide a detailed explicit rationale for them.

It is not difficult to construct situations in which people's intuitions will be misleading. This is especially true of situations that are described with words that have technical meanings that differ from the common meanings of those words as they are used in everyday speech.

Lack of understanding of basic concepts such as randomness or event independence can support the gambler's fallacy in its various forms and related misconceptions. Failure to make a clear distinction between specific events and events with specific properties can cause confusion about event probabilities.

In many situations, people tend to extract less information from observations than there is to be extracted, according to Bayes's rule, and modify existing beliefs less than they should in the light of new evidence. On the other hand, they also sometimes appear to attach too much weight to case-specific information and to discount the importance of base rates. These results, both of which have much experimental support, are, at least on the surface, somewhat inconsistent with each other and remain to be reconciled fully.

What role base rates should play in probabilistic reasoning, and how to determine what the appropriate base rates are in any given situation are difficult questions. Some of the work on base-rate neglect has not been sufficiently sensitive to this fact and has proceeded as though the answers were obvious and unequivocal, but the controversy surrounding the issue makes it clear that they are not.

When probabilistic information is acquired piecemeal over a period of time, the decisions that people make on the basis of that information, or the conclusions they draw from it, are not independent of the order in which the information is obtained. The information that is acquired early can contribute to the shaping of an initial tentative conclusion that then influences the interpretation of the information that is subsequently acquired.

Precisely how alternatives are worded, or framed, can influence the selections people make when asked to choose between events that can be specified only in probabilistic terms. Some of this influence may be attributable to the fact that certain expressions are commonly construed to have a meaning in everyday discourse that differs from the meanings intended in the experimental situations. Beliefs about the probable outcomes of uncertain events can also be influenced by illusory assumptions of personal control over chance events. Memorable personal experiences or vivid accounts of the personal experiences of others can easily have greater influence on people's beliefs about probabilistic events than dry statistical summaries of large bodies of data.

People have some appropriate intuitions about sample size—an approximate notion of the law of large numbers, for example—but these intuitions are not very precise and they appear to be overridden easily in specific instances.

Predictions about probabilistic events also can be influenced by people's preferences for some events over others.

Estimations of covariation or contingency are often distorted by a propensity to focus on instances of co-occurrence of the events of interest and to neglect to give adequate consideration to those occasions on which one of the events occurs without the other. The tendency to see correlations where they do not exist or to overestimate the strength of a contingency relationship has been recognized in the phenomenon of illusory correlation.

People often overestimate the probability of the conjunction of two or more events and underestimate the probability of event disjunctions. Estimates of the probability of the conjunction of two or more events sometimes violate the principle that the conjunction of events cannot be greater than the probability of the least probable component event. This conjunction fallacy is especially likely to occur when the conjunctive event in question is more representative of the conjunctive set than of the smaller of the component sets.

People appear to rely on a variety of heuristic rules for reasoning about probabilistic events. These rules appear to work quite well in many of the probabilistic reasoning situations that people face in everyday life. It has been possible to design situations in the laboratory in which the same heuristics lead people to respond in ways that would be considered irrational if viewed in isolation. The question of whether the use of these heuristics is rational when viewed in the larger context that takes one's total life experience and human cognitive limitations into account is a matter of debate.

People's limitations in dealing with probabilistic situations appear to derive from several sources including lack of knowledge of certain concepts and principles that are basic to probability theory, computational or capacity limitations, overreliance on simplified rules of thumb that work well in some but not all contexts, and motivational failures that result in people underperforming their capabilities because of lack of effort.

Some of these limitations would seem to be addressable by education and there is evidence that training in probability theory and statistics can improve the quality of reasoning under uncertainty, especially if that training is designed to exploit and sharpen the intuitions about probability and statistics that people appear to have even in the absence of training.

To the extent poor reasoning under uncertainty that has been observed in the laboratory is a motivational problem reflecting an unwillingness to make the effort necessary to reason well, education is likely to be only a partial answer. And any educational process that fosters the idea that people can reason well without effort, about either probabilistic events or anything else, is helping to perpetuate an unfortunate myth.

CHAPTER 12

Concluding Comments

Gambling, the drawing of lots, and other activities that today we associate with chance go back to antiquity. Whether the ancients had a concept of chance that was very close to what is generally meant by chance today is questionable. The use of chance devices for purposes of divination suggests that users, at least sometimes, believed the outcomes to be revealing of otherwise hidden truths: Was the accused party guilty or not? Would the king's army prevail in battle? ...

As a mathematical discipline, probability theory is a relative late-comer; historians of mathematics usually mark its beginning in the middle of the 17th century, with the work of Blaise Pascal, Pierre de Fremat, and some of their contemporaries. It was axiomatized only early in the 20th century as a consequence of the work, primarily, of Andrei Kolmogorov. Today well-formed probability problems generally can be solved straightforwardly with the application of the probability calculus, as laid out, for example, by William Feller (1957, 1966). The usefulness of probability theory has been established beyond doubt in numerous contexts and its impact on the sciences, especially during the 20th century, has been profound.

Despite these facts, the philosophical question of what probability and closely related entities like chance and randomness really are remains a matter of debate. All these terms have been given a variety of connotations, so being explicit about what is intended in specific contexts can help avoid confusion and misunderstanding. Whatever chance is, it is not the antithesis of lawfulness, as the predictability of chance events in the aggregate attests. Probability paradoxes and dilemmas of various sorts provide interesting opportunities to study probabilistic reasoning, and have sometimes played important roles in the development of theories of choice behavior under uncertainty.

Statistics, which began as a strictly descriptive discipline—a means of dealing with variability, especially in demographic data—evolved over time into an indispensable tool for all the sciences, hard and soft. It is widely used for both descriptive and inferential purposes and, like any powerful tool, is subject to misuses and abuses of various sorts, either by intent or from an inadequate understanding of the assumptions on which specific statistical procedures are based.

How good are people's untutored intuitions about probability, chance, randomness, and closely related concepts? Certainly they are far from perfect. There are numerous examples in the literature of probabilistic intuitions of even prominent mathematicians that have proved to be wrong. And the evidence that people in general often misjudge probabilities, sometimes systematically, is strong. A variety of errors of judgment involving probabilistic situations that appear to be commonly made have been identified; the gambler's fallacy in various guises, illusory correlations, the conjunction fallacy, overconfidence in one's own judgments on probabilistic matters, and the confirmation bias are cases in point. Others could be mentioned. Confusions involving conditional probabilities—failure to distinguish correctly what is conditional on what—have often been reported.

It is also the case, however, that some experimental results that have been widely taken as evidence of faulty probabilistic reasoning can be interpreted in other ways. Probability problems can very easily be stated incompletely or ambiguously. The incompleteness or ambiguousness may go unnoticed, and different people may unwittingly make different assumptions about the statement's intended meaning—and arrive at different problem solutions.

Research on how people make choices under uncertainly has revealed frequent departures from the predictions, or prescriptions, of normative models that assume that choices are driven by the objective of maximizing expected utility. Behavioral data have led to the development of alternative accounts of how people make choices under uncertainty that emphasize the use of specific heuristic strategies that work effectively in many contexts but that yield undesirable consequences in others; precisely what those heuristic strategies are and the conditions under which they are effective are matters of continuing research. An unresolved issue is the extent to which many of the findings with choice tasks in the laboratory generalize to real-world situations of interest.

In sum, as intuitive statisticians, people are neither at a total loss nor remarkably facile. They evidence some sound intuitions even without training, but they also typically display a variety of misconceptions about probabilistic variables, and problems requiring a more than superficial understanding of probability theory are likely to cause significant difficulties. Training in probability theory and mathematical statistics tends to improve probabilistic reasoning—it would be surprising and disheartening indeed if this were not so—but there are probability problems that can be difficult even for experts, who sometimes disagree regarding what the solutions are. In the final analysis, we are all pushed back to our intuitions, informed by training or not; we can accept only what we find plausible. But a motivating premise of this book is that plausibility that rests on familiarity with the development of probabilistic thinking over the centuries, with various perspectives that have been offered in recent times, and with a body of experimental work on the topic is to be preferred to plausibility that rests on uninformed intuition alone.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Abelson, R. P. (1997a). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.

Abelson, R. P. (1997b). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.) *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Lawrence Erlbaum Associates.

Adams, P. A., & Adams, J. K. (1960). Confidence in the recognition and reproduction of words difficult to spell. *American Journal of Psychology, 73,* 544–552.

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on predication. *Journal of Personality and Social Psychology, 35,* 303–314.

Alfidi, J. (1971). Informed consent: A study of patient reaction. *Journal of the American Medical Association, 216,* 1325–1329.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621–1630.

Allais, M. (1990). Criticism of the postulates and axioms of the American School. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 113–139). New York: Cambridge University Press. (Original work published 1979)

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology, 34,* 1–11.

Allen, F. W. (1987). Towards a holistic appreciation of risk: The challenges for communicators and policy makers. *Science, Technology, and Human Values, 12,* 138–143.

Alloy, L. B. (1988). Expectations and situational information as cocontributors to covariation assessment: A reply to Goddard and Allan. *Psychological Review, 95,* 299–301.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: Genenal, 108,* 441–485.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and by animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

Allwood, C. M., & Granhag, P. A. (1996). Considering the knowledge you have: Effects on realism in confidence judgements. *European Journal of Cognitive Psychology, 8,* 235–256.

Alpert, M, & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press.

Alvarez, L. W. (1965). A pseudo experience in parapsychology [Letter to the Editor]. *Science, 148,* 1541.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, N. H. (1964). Test of a model for number-averaging behavior. *Psychonomic Science, 1,* 189–190.

Anderson, N. H. (1979). Algebraic rules in psychological measurement. *American Scientist, 67*(5), 555–563.

Apostolakis, G. (1990). The concept of probability in safety assessments of technological systems. *Science, 250,* 1359–1364.

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society, 27,* 186–190.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39,* 133–144.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37,* 93–110.

Arkes, H. R., & Hammond, K. R. (Eds.). (1986). *Judgment and decision making: An interdisciplinary reader.* New York: Cambridge University Press.

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General, 112,* 117–135.

Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New York: Wiley.

Arrow, K. J. (1971). *Essays in the theory of risk-bearing.* Chicago: Markham.

Arrow, K. J. (1990). Values and collective decision making. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 337–353). New York: Cambridge University Press.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46,* 81–86.

Axelrod, R. (1980a). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution, 24,* 3–25.

Axelrod, R. (1980b). More effective choice in the prisoner's dilemma. *Journal of Conflict Resolution, 24,* 379–403.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211,* 1390–1396.

Ayala, F. J. (1978). The mechanisms of evolution. *Scientific American, 239*(3), 56–69.

Ayer, A. J. (1965). Chance. *Scientific American, 213*(4), 44–54.

Ayton, P., Hunt, A. J., & Wright, G. (1991a). Psychological conceptions of randomness. *Journal of Behavioral Decision Making, 2,* 221–238.

Ayton, P., Hunt, A. J., & Wright, G. (1991b). Randomness and reality, Reply to comments on "Psychological conceptions of randomness." *Journal of Behavioral Decision Making, 2,* 222–226..

Babad, E., & Katz, Y. (1992). Wishful thinking—Against all odds. *Journal of Applied Social Psychology, 21,* 1921–1938.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 1–29.

Baker, J. D., McKendry, J. M., & Mace, D. J. (1968). Certitude judgments in an operational environment (Tech. Res. Note No. 200). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ball, W. W. R., & Coxeter, H. S. M. (1987). *Mathematical recreations and essays* (13th ed.). New York: Dover. (First edition authored by Ball and published 1892)

Barber, B. (1961). Resistence by scientists to scientific discovery. *Science, 134,* 596–602.

Bar–Hillel, M. A. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance, 9,* 396–406.

Bar-Hillel, M. A. (1974). Similarity and probability. *Organizational Behavior and Human Performance, 11,* 277–282.

Bar-Hillel, M. A. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance, 24,* 245–257.

Bar-Hillel, M. A. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Bar-Hillel, M. A. (1983). The base-rate fallacy controversy. In R. W. Scholtz (Ed.), *Decision making under uncertainty.* Amsterdam: North Holland.

Bar-Hillel, M. A. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica, 55,* 91–107.

Bar-Hillel, M. A., & Budescu, D. (1995). The elusive wishful-thinking effect. *Thinking and Reasoning, 1,* 71–103.

Bar-Hillel, M. A., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11,* 109–122.

Bar-Hillel, M., & Fischhoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology, 41,* 671–680.

Bar-Hillel, M. A., & Wagenaar, W. A. (1993). The perception of randomness. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 369–393). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baril, G. L., & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist, 50,* 1098–1099.

Baron, J. (1985). *Rationality and intelligence.* New York: Cambridge University Press.

Baron, J. (1988). *Thinking and deciding.* New York: Cambridge University Press.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making.* New York: Oxford University Press.

Barrow, J. D. (1990). *The world within the world.* New York: Oxford University Press.

Barrow, J. D. (1991). *Theories of everything: The quest for ultimate explanation.* New York: Oxford University Press.

Barrow, J. D. (1992). *Pi in the sky: Counting, thinking, and being.* New York: Oxford University Press.

Barrow, J. D. (1998). *Impossibility: The limits of science and the science of limits.* New York: Oxford University Press.

Barrow, J. D., & Silk, J. (1983). *The left hand of creation.* New York: Basic Books.

Barrow, J. D., & Tipler, F. J. (1988). *The anthropic cosmological principle.* New York: Oxford University Press.

Bartecchi, C. E., MacKenzie, T. D., & Schrier, R. W. (1995). The global tobacco epidemic. *Scientific American, 272*(5), 44–51.

Baum, A., Fleming, R., & Singer, J. E. (1983). Coping with victimization by technological disaster. *Journal of Social Issues, 39,* 117–138.

Bauman, L. J., & Siegel, J. (1987). Misperception among gay men of the risk of AIDS. *Journal of Applied Social Psychology, 17,* 329–350. Berger (1997).

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, 53,* 370–418. (Reprinted in G. A. Barnard (1958), Studies in the history of probability and statistics. *Biometrika, 45,* 293–315)

Beach, L. R. (1966). Accuracy and consistency in the revision of subjective probabilities. *IEEE Transactions in Human Factors in Electronics HFE-7,* 29–37.

Beach, L. R., & Peterson, C. R. (1966). Subjective probabilities for unions of events. *Psychonomic Science, 5,* 307–308.

Beach, L. R., & Scopp, T. S. (1966). Inferences about correlations. *Psychonomic Science, 6,* 253–254.

Beach, L. R., & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance, 3,* 109–123.

Beach, L. R., Smith, B., Lundell, J., & Mitchell, T. R. (1988). Image theory: Descriptive sufficiency of a simple rule for the compatibility test. *Journal of Behavioral Decision Making, 1,* 17–28.

Beach, L. R., & Swensson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science, 5,* 161–162.

Beatty, J., Cartwright, N., Coleman, W., Gigerenzer, G., & Morgan, M. S. (1987). Introduction to volume 2. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 1–4). Cambridge, MA: MIT Press.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science, 8,* 41–55.

Bell, E. T. (1937). *Men of mathematics: The lives and achievements of the great mathematicians from Zeno to Poincaré.* New York: Dover.

Bell, P. A., Fisher, J. D., Baum, A., & Greene, T. C. (1990). *Environmental psychology.* Fort Worth, TX: Holt, Rinehart & Winston.

Beltrami, E. (1999). *What is random? Chance and order in mathematics and life.* New York: Springer-Verlag.

Bem, D. J., Wallach, M. A., & Kogan, N. (1965). Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology, 1,* 453–460.

Bennett, D. J. (1998). *Randomness.* Cambridge, MA: Harvard University Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust reciprocity and social history. *Games and Economic Behavior, 10,* 122–142.

Berg, J., Dickhaut, J., & O'Brien, J. (1985). Preference reversal and arbitrage. In V. Smith (Ed.), *Research in experimental economics* (pp. 31–72). Greenwich, CT: JAI.

Berger, J. O. (1997). *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76,* 159–165.

Berkeley, D., & Humphreys, P. (1982). Structuring decision problems and the "bias heuristic." *Acta Psychologica, 50,* 201–252.

Berl, J., Lewis, G., & Morrison, R. S. (1976). Applying models of choice to the problem of college selection. In J. S. Carroll & J. W. Psyne (Eds.), *Cognition and social behavior* (pp. 203–220). Hillsdale, NJ: Lawrence Erlbaum Associates.

Berlinski, D. (2000). *Newton's gift: How Sir Isaac Newton unlocked the system of the world.* New York: Simon & Schuster.

Bernoulli, D. (1738). Specimen theoriae norae de mensura sortis. *Commentarii Academoae, Scientiarum Imperiales Petropolitanae, 5,* 175–192. (Translation published in 1954 by L. Sommer in *Econometrica, 22,* 23–26)

Bernoulli, J. (1713). *Ars conjectandi [The art of conjecture].* Basel, Switzerland:

Bernstein, P. L. (1996). *Against the gods: The remarkable story of risk.* New York: Wiley.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination: Using few cues to choose. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 235–254). New York: Oxford University Press.

Berry, R. J. (1988). *God and evolution: Creation, evolution and the Bible.* London: Hodder & Stoughton.

Bertrand, J. (1889). *Calcul des probabilities [Calculus of probabilities].* Paris: Gautier-Villars et Fils.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory and Cognition, 10,* 511–519.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187,* 398–404.

Billings, R. S., & Schaalman, M. L. (1980). Administrators' estimation of the probability of outcomes of school desegregation: A field test of the availability heuristic. *Organizational Behavior and Human Performance, 26,* 97–114.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70,* 101–109.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96,* 85–94.

Black, D. (1958). *The theory of committees and elections.* Cambridge, England: Cambridge University Press.

Blackburn, S. (1981). Rational animal? Commentary on Cohen, 1981. *The Behavioral and Brain Sciences, 4,* 331–332.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM, 28,* 289–299.

Blyth, C. R. (1972a). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Society, 67,* 364–366.

Blyth, C. R. (1972b). Rejoinder. *Journal of the American Statistical Society, 67,* 379–381.

Blyth, C. (1972c). Some probability paradoxes in choice among random alternatives. *Journal of the American Statistical Association, 67,* 366–373.

Blythe, P. W., Todd, P. M., & Miller, G. F. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 257–285). New York: Oxford University Press.

Bolin, B., & Doos, B. R. (1986). *The greenhouse effect: Climatic change and ecosystems.* New York: Wiley.

Boorstin, D. J. (1985). *The discoverers: A history of man's search to know his world and himself.* New York: Vintage Books.

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 59–72). New York: Oxford University Press.

Borgida, E., & Brekke, N. (1981). *New directions in attribution research* (Vol. 5). Hillsdale, NJ: Lawrence Erlbaum Associates.

Borgida, E., & Howard-Pitney, B. (1983). Personal involvement and the robustness of perceptual salience effects. *Journal of Personality and Social Psychology, 45,* 560–570.

Bower, G. H. (1995). *Emotion and social judgments.* Washington, DC: Federation of Behavioral, Psychological and Cognitive Sciences.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist.* New York: Wiley.

Boyd, R. (1988). Is the repeated prisoner's dilemma a good model of reciprocal altruism? *Ethology and Sociobiology, 9,* 211–222.

Bradbury, J. A. (1989). The policy implications of differing concepts of risk. *Science, Technology, and Human Values, 14,* 380–399.

Bradley, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society, 17,* 82–84.

Braithwaite, R. B. (1974). The predictionist justification of induction. In R. Swinburne (Ed.), *The justification of induction* (pp. 102–126). London: Oxford University Press. (Original work published 1953)

Brams, S. J. (1976). *Paradoxes in politics.* New York: The Free Press.

Brams, S. J., & Fishburn, P. C. (1983). *Approval voting.* Boston: Birkhauser.

Brekke, N., & Borgida, E. (1988). Expert psychological testimony in rape trials: A social-cognitive analysis. *Journal of Personality and Social Psychology, 55,* 372–386.

Brenner, M. W. (1973). The next-in-line effect. *Journal of Verbal Learning and Verbal Behavior, 12,* 320–323.

Brenner, M. W. (1976). *Memory and interpersonal relations.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor, MI.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10,* 252–268.

Brewer, J. K., & Owen, P. W. (1973). A note on the power of statistical tests in the "Journal of Educational Measurement." *Journal of Educational Measurement, 10,* 71–74.

Brody, C. J. (1984). Differences by sex in support for nuclear power. *Social Forces, 63,* 209–228.

Brown, R. V., Kahr, A. S., & Peterson, C. R. (1974). *Decision analysis for the manager.* New York: Holt, Rinehart & Winston.

Browne, M. W., & Cudeck, R. C. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21,* 230–258.

Brubaker, E. R. (1975). Free ride, free revelation, or the golden rule? *Journal of Law and Economics, 18,* 147–161.

Buchanan, J. M. (1978). *Cost and choice: An inquiry in economic theory.* Chicago: University of Chicago Press.

Buckle, H. T. (1857–1861). *History of civilization in England* (2 vols.). London: J. W. Parker.

Budescu, D. V. (1985). Analysis of dichotomous variables in the presence of serial dependence. *Psychological Bulletin, 97,* 547–561.

Budescu, D. V., & Bruderman, M. (1995). The relationship between the illusion of control and the desirability bias. *Journal of Behavioral Decision Making, 8,* 109–125.

Buehler, R., Griffin, D., & McDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin, 23,* 238–247.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67,* 366–381.

Buehler, R., Griffin, D., & Ross, M. (1995). It's about time: Optimistic predictions in love and work. In W. Stroebe & M. Hewstone (Eds.), *The European Review of Social Psychology* (Vol. 6, pp. 1–32). Chichester, England: Wiley.

Buffon, G. L. L. (1777). Essai d'arithmetique morale [Essay on moral arithmetic]. *Supplément à l'histoire naturelle* (Vol. 44).

Bunge, M. (1996). In praise of intolerance to charlatanism in academia. In P. R. Gross, N. Levitt, & M. W. Lewis (Eds.), *The flight from reason* (pp. 96–115). New York: New York Academy of Sciences.

Burke, J. G. (1966). Bursting boilers and the federal power. *Technology and Culture, 7,* 1–23.

Burton, I., Kates, R. W., & White, G. R. (1978). *The environment as hazard.* New York: Oxford University Press.

Busemeyer, J. R. (1990). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory* (Vol. 1, pp. 187–215). Hillsdale, NJ: Lawrence Erlbaum Associates.

Butler, C. (1970). *Number symbolism*. London: Routledge & Kegan Paul.

Byram, S. J. (1997). Cognitive and motivational factors influencing time prediction. *Journal of Experimental Psychology: Applied, 3,* 216–239.

Camerer, C. (1988). Illusory correlations in perceptions and predictions of organizational traits. *Journal of Behavioral Decision Making, 1,* 77–94.

Camerer, C. (1990). Do markets correct biases in probability judgment? Evidence from market experiments. In L. Green & J. H. Kagel (Eds.), *Advances in behavioral economics* (Vol. 2, pp. 126–172). Norwood, NJ: Ablex.

Camerer, C., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives, 9,* 209–219.

Carlson, B. W. (1990). Anchoring and adjustment in judgments under risk. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 655–676.

Carnap, R. (1953). What is probability? *Scientific American, 189*(3), 128–138.

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1950)

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior, 10,* 722–729.

Carter, B. (1974). Large number coincidences and the anthropic principle in cosmology. In M. S. Longair (Ed.), *Confrontation of cosmological theories with observational data* (pp. 291–298). Dordrecht, Netherlands: Reidel.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287–292.

Cassells, W., Schoenberger, A., & Grayboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999.

Catania, A. C., & Cutts, D. (1963). Experimental control of superstitious responding in humans. *Journal of Experimental Analysis of Behavior, 6,* 203–208.

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology, 50,* 492–501.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use fo source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39,* 752–766.

Chaitin, G. J. (1975). Randomness and mathematical proof. *Scientific American, 232*(5), 47–52.

Chapman, G. B., & Bornstein, B. H. (1996). The more you ask for, the more you get: Anchoring in personal injury verdicts. *Applied Cognitive Psychology, 10,* 519–540.

Chapman, G, B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7,* 223–242.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6,* 151–155.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280.

Chenzoff, A., Crittenden, R. L., Flores, I., & Tolcott, M. A. (1960). *Human decision making as related to air surveillance systems: A survey of literature and current research.* (Tech Rep. No. 1, AFCCDD-TR-60-25). U.S. Air Force.

Chow, A., Crittenden, R. L., Flores, I., & Tolcott, M. A. (1960). *Human decision making as related to air surveillance systems: A survey of literature and current research.* (Tech. Rep. No. 1, AFCCDD-TR-60-25). U.S. Air Force

Chow, S. L. (1987). *Experimental psychology: Rationale, procedures and issues.* Calgary, Alberta, Canada: Detselig Enterprises.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

Chow, S. L. (1989). Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin, 106,* 161–165.

Chow, S. L. (1991). Conceptual rigor versus practical impact. *Theory and Psychology, 1,* 337–360.

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility.* Beverly Hills, CA: Sage.

Chow, S. L. (1998a). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21,* 169–239 (including commentary).

Chow, S. L. (1998b). What statistical significance means. *Theory and Psychology, 8,* 323–330.

Christensen-Szalanski, J. J., & Beach, L. R. (1982). Experience and the base rate fallacy. *Organizational Behavior and Human Performance, 29,* 270–278.

Christensen-Szalanski, J. J. J., Beck, D. E., Christensen-Szalanski, C. M., & Keopsell, T, D. (1983). Effects of expertise and experience on risk judgments. *Journal of Applied Cognition, 68,* 278–283.

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928–935.

Chronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language.* New York: Harcourt Brace.

Clark, R. D. (1971). Group-induced shift toward risk: A critical appraisal. *Psychological Bulletin, 76,* 251–270.

Clarke, R. D. (1946). An application of the Poisson distribution. *Journal of the Institute of Actuaries (London), 72,* 72.

Cohen, B., & Lee, I. (1979). A catalog of risks. *Health Physics, 36,* 707–722.

Cohen, I. B. (1987). Scientific revolutions, revolutions in science, and a probabilistic revolution 1800–1930. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 23–44). Cambridge, MA: MIT Press.

Cohen, J. (1957). Subjective probability. *Scientific American, 197*(5), 128–138.

Cohen, J. (1972). *Psychological probability: Or the art of doubt.* London: George Allen & Unwin.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cohen, J., Dearnaley, E. J., & Hansel, C. E. M. (1958). Skill with chance: Variations in estimates of skill with an increasing element of chance. *British Journal of Psychology, 49,* 319–323.

Cohen, L. J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, L. J. (1970). *The implications of induction.* London: Methuen.

Cohen, L. J. (1977). The probable and the provable. Oxford, England: Clarendon.

Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition, 7,* 385–407.

Cohen, L. J. (1980). Whose is the fallacy? A rejoinder to Daniel Kahneman and Amos Tversky. *Cognition, 8,* 89–92.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–331.

Cohen, L. J. (1982). Are people programmed to commit fallacies? Further thoughts about the interpretation of data on judgment. *Journal of the Theory of Social Behavior, 12,* 251–274.

Cohen, L. J., & Chesnick, E. I. (1970). The doctrine of psychological chances. *British Journal of Psychology, 61,* 323–334.

Cohen, L. J., Chesnick, E. I., & Haran, D. (1971). Evaluation of compound probabilities in sequential choice. *Nature, 232,* 214–216.

Cohen, L. J., Chesnick, E. I., & Haran, D. (1972). A confirmation of the inertial—y—effect of sequential choice and decision. *British Journal of Psychology, 63,* 41–46.

Cohen, M. D., & Marsh, J. G. (1974). *Leadership and ambiguity: The American College President.* New York: McGraw-Hill.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method.* London: Routledge & Kegan Paul.

Cole, K. C. (1997). *The universe and the teacup: The mathematics of truth and beauty.* New York: Harcourt Brace.

Coleman, W. (1987). Experimental physiology and statistical inference: The therapeutic trial in nineteenth-century Germany. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 201–226). Cambridge, MA: MIT Press.

College Board. (1976–1977). *Student descriptive questionnaire.* Princeton, NJ: Educational Testing Service.

Combs, B., & Slovic, P. (1979). Causes of death: Biased newspaper coverage and biased judgments. *Journalism Quarterly, 56,* 837–843.

Compendium of curious coincidences: Parallels in the lives and deaths of A. Lincoln and J. F. Kennedy. (1964, August 21). *Time, 84,* 19.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues* la pluralité des voix [Essay on the application of analysis of the probability of decision of votes rendered by a plurality]. Paris: De l'Imprimerie Royale.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Coombs, C. H., Bezembinder, T. G., & Goode, F. M. (1967). Testing expectation theories of decision making without utility or subjective probability. *Journal of Mathematical Psychology, 4,* 72–103.

Cooper, A. C., Woo, C. Y., & Dunkelberg, W. C. (1988). Entrepreneurs' perceived chances for success. *Journal of Business Venturing, 3,* 97–108.

Copi, I. M. (1968). *Introduction to logic* (3rd ed.). New York: Macmillan.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161–172.

Cosmides, L., & Tooby, J. (1990, August). *Is the mind a frequentist?* Paper presented at the second annual meeting of the Human Behavior and Evolution Society, Los Angeles.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités [Exposition of the theory of chances and probabilities].* Paris: Hachette.

Covello, V. T., & Johnson, B. B. (1987). The social and cultural construction of risk: Issues, methods, and case studies. In B. B. Johnson & V. T. Covello (Eds.), *The social and cultural construction of risk* (pp. vii–xiii). Dordrecht, Netherlands: D. Reidel.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4,* 49–70.

Cramer, G. (1728). Letter to N. Bernoulli, published by D. Bernoulli (1738).

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292.

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin, 8,* 214–220.

Crouch, E. A. C., & Wilson, R. (1982). *Risk/benefit analysis.* Cambridge, MA: Ballinger.

Crowe, B. (1969). The tragedy of the commons revisited. *Science, 166,* 1103–1107.

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory and Cognition, 22,* 249–257.

Cvetkovich, G., & Earle, T. C. (Eds.). (1992). Public responses to environmental hazards. *Journal of Social Issues, 48*(4), 1–187.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York: Oxford University Press.

Dale, A. I. (1974). On a problem in conditional probability. *Philosophy of Science, 41,* 204–206.

Dale, H. C. A. (1960). A study of subjective probability. *The British Journal of Statistical Psychology, 13 (Pt. 1),* 19–29.

Dale, H. C. A. (1968). Weighing evidence: An attempt to assess the efficiency of the human operator. *Ergonomics, 11,* 215–230.

Damas, P. A., Goodman, B. C., & Peterson, C. R. (1972). *Bayes theorem: Response scales and feedback* (University of Michigan Project No. 197–014). Ann Arbor, MI.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8,* 377–383.

Daston, L. J. (1987a). The domestication of risk: Mathematical probability and insurance 1650–1830. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 238–260). Cambridge, MA: MIT Press.

Daston, L. J. (1987b). Rational individuals versus laws of society: From probability to statistics. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 295–304). Cambridge, MA: MIT Press.

Daston, L. J. (1988). *Classical probability in the enlightenment.* Princeton, NJ: Princeton University Press.

David, F. N. (1962). *Games, gods, and gambling: The origins and history of probability.* New York: Hafner.

Davies, P. C. W. (1982). *The accidental universe.* New York: Cambridge University Press.

Davies, P. C. W. (1983). *God and the new physics.* New York: Simon & Schuster.

Davies, P. (1988). *The cosmic blueprint: New discoveries in nature's creative ability to order the universe.* New York: Simon & Schuster.

Davies, P. C. W. (1992). *The mind of God: The scientific basis for a rational world.* New York: Simon & Schuster.

Davis, H. L., Hoch, S. J., & Ragsdale, E. E. (1986).An anchoring and adjustment model of spousal predictions. *Journal of Consumer Research, 13,* 25–37.

Davis, J. N., & Todd, P. M. (1999). Parental investment by simple decision rules. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 309–324). New York: Oxford University Press.

Davis, P. J., & Hersh, R. (1981). *The mathematical experience.* Boston: Houghton Mifflin.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188.

Dawes, R. (1976). Shallow psychology. In J. S. Carroll & J. W. Payne (Eds.). *Cognition and social behavior* (pp. 3–12). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* New York: Harcourt Brace.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision-making. *Psychological Bulletin, 81,* 95–106.

de Dombal, F. T., Leaper, D. J., Horrocks, J. C., Staniland, J. R., & McCann, A. P. (1974). Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians. *British Medical Journal, 1,* 376–380.

de Finetti, B. (1962). Does it make sense to speak of "good probability appraiser?" In I. J. Good, A. J. Mayne, & J. M. Smith (Eds.), *The scientist speculates: An anthology of half-baked ideas* (pp. 357–364). New York: Basic Books.

de Finetti, B. (1976). Probability: Beware of falsifications! *Scientia, 3,* 283–303.

DeGroot, M. H. (1982). Comment (on Schafer [1982]). Lindley's paradox. *Journal of the American Statistical Association, 77,* 325–334). *Journal of the American Statistical Association, 77,* 336–339.

DeJoy, D. (1989). The optimism bias and traffic accident risk prevention. *Accident Analysis and Prevention, 21,* 333–340.

De Moivre, A. (1756). Excerpt from *The doctrine of chances.* Reprinted as Appendix 5 in F. N. David (1962). *Games, gods, and gambling: The origins and history of probability* (pp. 254–267). New York: Hafner.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66,* 819–829.

Dennett, D. C. (1981). Three kinds of intentional psychology. In R. Healey (Ed.), *Reduction, time and reality* (pp. 37–61). Cambridge, England: Cambridge University Press.

Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58,* 952–963.

Devine, P. G., & Ostrom, T. M. (1985). Cognitive mediation of inconsistency discounting. *Journal of Personality and Social Psychology, 49,* 5–21.

Devlin, K. (2000a). *The language of mathematics.* New York: Freeman.

Devlin, K. (2000b). Snake eyes in the Garden of Eden. *The Sciences, 40*(4), 14–17.

Dewdney, A. K. (1993). *200 percent of nothing.* New York: Wiley.

Diaconis, P., & Freedman, D. (1981). The persistence of cognitive illusions. Commentary on Cohen, 1981. *The Behavioral and Brain Sciences, 4,* 333–334.

Diaconis, P., & Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association, 84,* 853–861.

Dickens, C. (1894). *The old curiosity shop.* Boston, MA: Houghton Mifflin. (Original work published 1840)

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attention. *Quarterly Journal of Experimental Psychology, 36A,* 29–50.

Dickinson, R. E., & Cicerone, R. J. (1986). Future global warming from atmospheric trace gases. *Nature, 319,* 109–115.

Ditto, P. H., Jemmott, J. B., & Darley, J. M. (1988). Appraising the threat of illness: A mental representational approach. *Health Psychology, 7,* 183–201.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63,* 568–584.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin and Review, 5,* 390–396.

Doherty, M. E., & Falgout, K. (1986). *Subjects' data selection strategies for assessing event covariation.* Unpublished manuscript, Department of Psychology, Bowling Green State University, Bowling Green, OH.

Donmell, M. L., & DuCharme, M. W . (1975). The effect of Bayesian feedback on learning in an odds estimation task. *Organizational Behavior and Human Performance, 14,* 305–313.

Dowie, J., & Elstein, A. (Eds.) (1988). *Professional judgment. A reader in clinical decision making.* Cambridge, England: Cambridge University Press.

Dracup, C. (1995). Hypothesis testing—what it really is. *The Psychologist, 8,* 359–362.

DuCharme, W. M., & Peterson, C. R. (1969). Proportion estimation as a function of proportion and sample size. *Journal of Experimental Psychology, 81,* 536–541.

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition, 9,* 85–110.

Dunning, D. (1993). Words to live by: The self and definitions of social concepts and categories. In J. Suls (Ed.), *Psychological perspectives on the self* (pp. 99–126). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dyson, F. J. (1971). Energy in the universe. *Scientific American, 225*(3), 50–59.

Eddington, A. S. (1935). *New pathways in science*. New York: Macmillan.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). New York: Cambridge University Press.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53,* 460–475, 644–663.

Edney, J. J. (1980). The commons problem: Alternative perspectives. *American Psychologist, 35,* 131–150.

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review, 69,* 109–135.

Edwards, W. (1965). Probabilistic information processing system for diagnosis and action selection. In J. Spiegel & D. Walker (Eds.), *Information system sciences. Proceedings of the second congress* (pp. 000–000). Washington, DC: Spartan Books.

Edwards, W. (1967). Dynamic decision theory and probabilistic information processing. *Human Factors, 4,* 59–73.

Edwards, W. (1968). Conservatism and human information processing. In B. Kleinmuntz (Ed.) *Formal representation of human judgment*. New York: Wiley.

Edwards, W., Lindman, H., & Phillips, L. D. (1965). Emerging technologies for making decisions. In *New directions in psychology II* (pp. 261–325). New York: Holt, Rinehart & Winston.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Edwards, W., & Tversky, A. (1967). *Decision making*. Middlesex, England: Penguin.

Einhorn, H. J. (1980). Overconfidence in judgment. In R.A. Shweder & D. W. Fiske (Eds.), *New directions for methodology of social and behavioral science: No. 4. Fallible judgment in behavioral research* (pp. 1–16). San Francisco: Jossey-Bass.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology, 32,* 53–88.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92,* 433–461.

Ekeland, I. (1993). *The broken dice*. Chicago: University of Chicago Press. (Original work published 1991 in French)

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics, 75,* 643–669.

Engle, P. L., & Lumpkin, J. B. (1992). How accurate are time-use reports? Effects of cognitive enhancement and cultural differences on recall accuracy. *Applied Cognitive Psychology, 6,* 141–159.

Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review, 62,* 693–752.

Erev, I., Bornstein, B. H., & Wallsten, T. S. (1993). The negative effect of probability assessment on decision quality. *Organizational Behavior and Human Decision Processes, 55,* 78–94.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 319–327.

Erikson, K. (1990, Fall). The fear you can't ignore. *Best of Business Quarterly,* pp. 52–59.

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology, 73,* 9–14.

Estes, W. K. (1964). *Probability learning.* In A. W. Melton (Ed.), *Categories of human learning* (pp. 90–128). New York: Academic Press.

Estes, W. K. (1976a). The cognitive side of probability learning. *Psychological Review, 83,* 37–64.

Estes, W. K. (1976b). Some functions of memory in probability learning and choice behavior. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10, pp. 2–45). New York: Academic Press.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Evans, J. St. B. T., & Bradshaw, H. (1986). Estimating sample-size requirements in research design: A study of intuitive statistical judgment. *Current Psychological Research and Reviews, 5*(1), 10–19.

Evans, J. St. B. T., & Dusoir, A. E. (1975). *Sample size and subjective probability judgements.* Paper presented at the meeting of the Experimental Psychology Society, Oxford, England.

Evans, J. St. B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica, 41,* 129–137.

Evans, J. St. B. P., & Pollard, P. (1985). Intuitive statistical inferences about normally distributed data. *Acta Psychologica, 60,* 57–71.

Faber, R. J. (1976). Re-encountering a counter-intuitive probability. *Philosophy of Science, 43,* 283–285.

Falk, R. (1979). Revision of probabilities and the time axis. In *Proceedings of the Third International Conference for the Psychology of Mathematics Education* (pp. 64–66). Warwick, England.

Falk, R. (1981). The perception of randomness. In *Proceedings of the Fifth International Conference for the Psychology of Mathematics Education* (pp. 222–229). Grenoble, France.

Falk, R. (1981–1982). On coincidences. *The Skeptical Inquirer, 6,* 18–31.

Falk, R. (1983). Experimental models for resolving probabilistic ambiguities. In R. Hershlowitz (Ed.), *Proceedings of the Seventh International Conference for the Psychology of Mathematics Education* (pp. 319–325). Rehovot, Israel: The Weizmann Institute of Science.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

Falk, R. (1989). Judgment of coincidences: Mine versus yours. *American Journal of Psychology, 102,* 477–493.

Falk, R. (1991). Randomness—An ill-defined but much needed concept. Commentary on "Psychological conceptions of randomness." *Journal of Behavioral Decision Making, 4,* 215–218.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 43,* 197–223.

Falk, R. (1993). *Understanding probability and statistics: A book of problems.* Wellesley, MA: A. K. Peters.

Falk, R., & Bar-Hillel, M. (1980). Magic possibilities of the weighted average. *Mathematics Magazine, 53,* 106–107.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75–98.

Falk, R., & Konold, C. (1992). The psychology of learning probability. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty-first century* (pp 151–164). Washington, DC: Mathematical Association of America.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104,* 301–318.

Falk, R., & MacGregor, D. (1983). The surprisingness of coincidences. In P. Humphreys, O. Svenson, & A. Vári (Eds.), *Analysing and aiding decision processes* (pp. 489–502). Budapest, Hungary: Adadémiai Kiadó.

Faust, D., Hart, K., & Guilmette, J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Counseling and Clinical Psychology, 56,* 578–582.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90,* 980–994.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica, 65,* 833–860.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics, 108,* 437–460.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114,* 817–868.

Feldman, J. M., Camburn, A., & Gatti, G. M. (1986). Shared distinctiveness as a source of illusory correlation in performance appraisal. *Organizational Behavior and Human Decision Processes, 37,* 34–59.

Feller, W. (1936–1937). Über das Gesetz der Grossen Zahlen [On the law of large numbers]. *Acta Litterarum ac Scientiarum Regiae Universitatis Hungaricae Francisco-Iosophinae, 8,* 191–201.

Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1, 2nd ed.). New York: Wiley.

Feller, W. (1966). *An introduction to probability theory and its applications, Vol. 2.* New York: Wiley.

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science, 4,* 282–296.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32–53.

Feynman, R. P. (1989). *The character of physical law.* Cambridge, MA: MIT Press. (Original work published 1965)

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123–129.

Fields, J. M., & Schuman, H. (1976). Public beliefs about the beliefs of the public. *Public Opinion Quarterly, 40,* 427–448.

Finke, R. A. (1984). Strategies for being random. *Bulletin of the Psychonomic Society, 22,* 40–41.

Fischer, G. W. (1982). Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organizational Behavior and Human Performance, 13,* 1–16.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 349–358.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, England: Cambridge University Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S. L., & Keeney, R. L. (1981). *Acceptable risk.* New York: Cambridge University Press.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1,* 155–172.

Fischhoff, B., & MacGregor, D. (1983). Judged lethality: How much people seem to know depends upon how they are asked. *Risk Analysis, 3,* 229–236.

Fischhoff, B., & Slovic, P. (1980). A little learning ... : Confidence in multicue judgment. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 779–800). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representations. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 330–344.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23,* 339–359.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1980). Knowing what you want; Measuring labile values. In T. S. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 117–141). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1982). Lay foibles and expert fables in judgments about risks. *The American Statistician, 36,* 241–255.

Fischhoff, B., Sverson, O., & Slovic, P. (1987). Active responses to environmental hazards: Perception and decision making. In D. Stokols & I Altman (Eds.), *Handbook of environmental psychology* (pp. 1089–1133). New York: Wiley.

Fishburn, P. C., & Kochenberger, G. A. (1979). Two-piece von Neumann–Morgenstern utility functions. *Decision Sciences, 10,* 503–518.

Fisher, A. (1988). *The logic of real arguments.* New York: Cambridge University Press.

Fischer, I., & Budescu, D. V. (1994). Desirability and hindsight biases in predicting results of a multi-party election. In J. P. Caverni, M. Bar-Hillel, H. Barron, & H. Jungermann (Eds.), *Frontiers in decision research* (pp. 193–211). Amsterdam: North Holland.

Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 181–197.

Flammarion, C. (1890). L'inconnu et les problèmes psychiques [The unknown and psychic problems]. Paris.

Fleming, R. A. (1970). The processing of conflicting information in a simulated tactical decision-making task. *Human Factors, 12,* 375–385.

Flexer, A. J., & Bower, G. H. (1975). Further evidence regarding instructional effects on frequency judgments. *Bulletin of the Psychonomic Society, 6,* 321–324.

Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin, 106,* 155–160.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Forgas, J. P., Bower, G. H., & Krantz, S. (1984). The influence of mood on perceptions of social interactions. *Journal of Personality and Social Psychology, 20,* 497–513.

Fourier, J. B. J. (1819). Extrait d'un mémoire sur la théorie analytique des assurances [Excerpt from a memoir on the analytical theory of assurances]. *Annales de Chimie et de Physique, 10,* 177–189.

Freudenthal, H. (1970). The aims of teaching probability. In L. Räde (Ed.), *The teaching of probability and statistics* (pp. 151–167). Stockholm: Almqvist & Wiksell.

Freund, J. E. (1965). Puzzle or paradox. *American Statistician, 19,* 29, 44.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Friedman, M. (1953). *Essays in positive economics*. Chicago: University of Chicago Press.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risks. *Journal of Political Economy, 56,* 279–304.

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1,* 149–157.

Galanter, E., & Pliner, P. (1974). Cross-modality matching of money against other continua. In H. R. Moskowitz, B. Scharf, & J. C. Stevens (Eds.), *Sensation and measurement* (pp. 65–76). Dordrecht, Netherlands: Reidel.

Galton, F. (1869). *Hereditary genius; An inquiry into its laws and consequences*. London: Macmillan.

Galton, F. (1874). *English men of science: Their nature and nurture*. London: Macmillan.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Galton, F. (1909). *Memories of my life* (3rd ed.). London: Methuen.

Gamow, G., & Stern, G. (1958). *Puzzle-math*. New York: Viking.

Gardenfors, P., & Sahlin, N.-E. (1983). Decision making with unreliable probabilities. *British Journal of Mathematical and Statistical Psychology, 36,* 240–251.

Gardner, M. (1957). *Fad and fallacies in the name of science*. New York: Dover.

Gardner, M. (1967). *The numerology of Dr. Matrix*. New York: Simon & Schuster.

Gardner, M. (1974). On the paradoxical situations that arise from nontransitive relations. *Scientific American, 231*(4), 110–114.

Gardner, M. (1976). Mathematical games: On the fabric of inductive logic, and some probability paradoxes. *Scientific American, 234*(3), 119–124.

Gardner, M. (1979). Mathematical games: In some patterns of numbers or words there may be less than meets the eye. *Scientific American, 241*(3), 22–32.

Gauthier, D. (1990). Maximization constrained: The rationality of cooperation. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 315–334). New York: Cambridge University Press. (Original work published 1986)

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology, 63,* 766–780.

Gell-Mann, M. (1994). *The quark and the jaguar: Adventures in the simple and the complex*. New York: Freeman.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gettys, C. F., Fisher, S. D., & Mehle, T. (1978). *Hypothesis generation and plausibility assessment* (Annual Rep. No. R 15-10-78. University of Oklahoma, Decision Processes Laboratory.

Gettys, C. F., & Wilke, T. A. (1969). The application of Bayes's theorem when the true data state is uncertain. *Organizational Behavior and Human Performance, 4,* 125–141.

Gibbs, W. W. (1994). Software's chronic crisis. *Scientific American, 271*(3), 86–95.

Gibson, B., Sanbonmatsu, D. M., & Posavac, S. S. (1997). The effects of selective hypothesis testing on gambling. *Journal of Experimental Psychology: Applied, 3,* 126–142.

Giere, R. N. (1972). The significance test controversy. *British Journal for the Philosophy of Science, 23,* 170–181.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1991b). How to make cognitive illusions disappear: Beyond "heuristics and biases." In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 83–115). London: Wiley.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 284–313). London: Routledge.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). New York: Wiley.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The Take the Best heuristic. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 75–95). New York: Oxford University Press.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gigerenzer, G. Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* New York: Cambridge University Press.

Gigerenzer, G., & Todd, P. M. (1999a). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (Eds.). (1999b). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal, 61,* 35–73.

Gilden, D. L., & Gray Wilson, S. (1995). Streaks in skilled performance. *Psychonomic Bulletin and Review, 2,* 260–265.

Gillies, D. A. (1973). *An objective theory of probability.* London: Methuen.

Gillman, L. (1992). The car and the goats. *American Mathematical Monthly, 99,* 3–7.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: The Free Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17,* 295–314.

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology, 52,* 464–474.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 702–718.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119–136.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory and Cognition, 10,* 597–602.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Glymour, C. (1996). Why I am not a Bayesian. In D. Papineau (Ed.), *Philosophy of science* (pp. 290–213). Oxford, England: Oxford University Press.

Goddard, M., & Allan, L. (1988). A critique of Alloy and Tabachnik's theoretical framework for understanding covariation assessment. *Psychological Review, 95,* 296–298.

Gold, E., & Hester, G. (1989). *The gambler's fallacy and the coin's memory.* Unpublished manuscript, Carnegie-Mellon University, Pittsburgh, PA.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender–Gestalt test. *Journal of Consulting Psychologists, 23,* 23–33.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from MMPI. *Psychological Monographs, 79* (Whole No. 9).

Goldberg, L. R. (1968). Simple models or simple processes? Some research in clinical judgment. *American Psychologist, 23,* 483–496.

Goldberg, S. (1976). Copi's conditional probability problem. *Philosophy of Science, 43,* 286–289.

Golding, S. L., & Rorer, L. G. (1972). Illusory correlation and subjective judgement. *Journal of Abnormal Psychology, 80,* 249–260.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge, MA: Harvard University Press.

Goldsmith, R. W. (1978). Assessing probabilities of compound events in a judicial context. *Scandinavian Journal of Psychology, 19,* 103–110.

Goldsteen, R., Schorr, J. K., & Goldsteen, K. S. (1989). Longitudinal study of appraisal at Three Mile Island: Implications for life event research. *Social Science and Medicine, 28,* 389–398.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

Goldstein, D., Hasher, L., & Stein, D. K. (1983). The processing of occurrence rate and item information by children of different ages and abilities. *American Journal of Psychology, 96,* 229–241.

Goldstein, W., & Einhorn, H. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94,* 236–254.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B, 14,* 107–114.

Good, I. J. (1972). Comment. *Journal of the American Statistical Association, 67,* 374–375.

Good I. J. (Ed) (1983a). *Good thinking: The foundations of probability and its applications.* Minneapolis: University of Minnesota Press.

Good, I. J. (1983b). The white shoe is a red herring. In I. J. Good (Ed.), *Good thinking: The foundations of probability and its applications* (pp.119–120). Minneapolis: University of Minnesota Press. (Original work published 1967)

Goodie, A. S., Ortmann, A., Davis, J. N., Bullock, S., & Werner, G. M. (1999). Demons versus heuristics in artificial intelligence, behavioral ecology, and economics. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 327–355). New York: Oxford University Press.

Gould, L. C., Gardner, G. T., DeLuca, D. R., Tiemann, A. R., Doob, L. W., & Stolwijk, J. A. J. (1988). *Perceptions of technological risks and benefits.* New York: Russell Sage Foundation.

Gould, S. J. (1989). The streak of streaks. *Chance: New Directions for Statistics and Computing, 2*(2), 10–16.

Gould, S. J. (1993). *Eight little piggies.* New York: Norton.

Graesser, A. C., & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory and Language, 30,* 186–209.

Graham, R. L., Rothschild, B. L., & Spencer, J. H. (1990). *Ramsey theory* (2nd ed.). New York: Wiley.

Graham, R. L., & Spencer, J. H. (1990). Ramsey theory. *Scientific American, 263*(1), 112–117.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69,* 54–61.

Green, C. H., & Brown, R. A. (1978). Counting lives. *Journal of Occupational Accidents, 2,* 55–70.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greening, L., & Chandler, C. C. (1997). Why it can't happen to me: The base rate matters, but overestimating skill leads to underestimating risk. *Journal of Applied Psychology, 27,* 760–780.

Greenstein, G. (1988). *The symbiotic universe: Life and the cosmos in unity.* New York: Morrow.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology, 43,* 89–99.

Grether, D. M. (1980). Bayes' rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics, 95,* 537–557.

Grether, D. M., & Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69,* 623–638.

Gribbin, J., & Rees, M. (1989). *Cosmic coincidences: Dark matter, mankind, and anthropic cosmology.* New York: Bantam.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. I. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York: Seminar Press.

Griffin, D. W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology, 59,* 1128–1139.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26,* 81–107.

Hacking, I (1965). *The logic of statistical inference.* Cambridge, England: Cambridge University Press.

Hacking, I. (1975). *The emergence of probability.* New York: Cambridge University Press.

Hacking, I. (1987a). Prussian numbers, 1860–1882. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 377–394). Cambridge, MA: MIT Press.

Hacking, I. (1987b). Was there a probabilistic revolution 1800–1930?. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 45–55). Cambridge, MA: MIT Press.

Hacking, I. (1990). *The taming of chance.* New York: Cambridge University Press.

Hacking, I. (2001). *An introduction to probability and inductive logic.* New York: Cambridge University Press.

Hake, H. W., & Hyman, R. (1953). Perception of the statistical structure of a random series of binary symbols. *Journal of Experimental Psychology, 45,* 64–74.

Hakohen, D., Avinon, O., & Falk, R. (1981). *What is surprising about coincidences?* Unpublished manuscript (in Hebrew), The Hebrew University, Jerusalem. (Cited in Falk, 1989)

Hallman, W. K., & Wandersman, A. H. (1992). Attribution of responsibility and individual and collective coping with environmental threats. *Journal of Social Issues, 48*(4), 101–118.

Hamil, R., Wilson, T. D., & Nisbett, R. E. (1980) Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology, 39,* 578–589.

Hamilton, D. L. (1979). A cognitive attributional analysis of stereotyping. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 53–84). New York: Academic Press.

Hamilton, D. L. (1981). Illusory correlation as a basis for stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 333–353). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hamilton, D. L., Dugan, P. M., & Trolier, T. K. (1985). The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 48,* 5–17.

Hamilton, D. L., & Gifford, R. (1976). Illusory correlation in interpersonal perception: A cognitive basis for stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392–407.

Hamilton, D. L., & Sherman, S. J. (1989). Illusory correlations: Implications for stereotype theory and research. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and prejudice: Changing conceptions* (pp. 59–82). New York: Springer-Verlag.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology, 101,* 252–254.

Hammond, K R. (2000). *Human judgment and social policy.* New York: Oxford University Press.

Hammond, K. R., & Marvin, B. A. (1981). *Report to the Rocky Flats Monitoring Committee concerning scientists' judgments of cancer risk* (Rep. No. 232). Boulder, CO: University of Colorado.

Hammond, K. R., McClelland, G. H., & Mumpower, J. (1980). *Human judgment and decision making.* New York: Hemisphere.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance, 9,* 30–34.

Hance, B. J., Chess, C., & Sandman, P. M. (1988). *Improving dialogue with communities: A risk communication manual for government.* Trenton, NJ: Department of Environmental Protection.

Hangenlocher, K. G. (1999). A Zeppelin for the 21st century. *Scientific American, 281*(5), 104–109.

Hansen, R. D., & Donoghue, J. M. (1977). The power of consensus: Information derived from one's own behavior. *Journal of Personality and Social Psychology, 35,* 294–302.

Hansen, W. B, & Malotte, C. K. (1986). Perceived personal immunity: The development of beliefs about susceptibility to the consequences of smoking. *Preventive Medicine, 15,* 363–372.

Hardin, G. (1968). The tragedy of the commons. *Science, 162,* 1243–1248.

Harkness, A. R., DeBono, K. G., & Borgida, E. (1985). Personal involvement and strategies for making contingency judgments: A stake in the dating game makes a difference. *Journal of Personality and Social Psychology, 49,* 22–32.

Harris, D. M., & Guten, S. (1979). Health protective behavior: An exploratory study. *Journal of Health and Social Behavior, 20,* 17–29.

Harris, J. M. (1981). The hazards of bedside Bayes. *Journal of American Medical Association, 246,* 2602–2605.

Harris, P. (1996). Sufficient grounds for optimism: The relationship between perceived controllability and optimistic bias. *Journal of Social and Clinical Psychology, 15,* 319–386.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8,* 8–11.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16,* 173–184.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388.

Hastie, R., Penrod, S. D., & Pennington, N. (1983). *Inside the jury.* Cambridge, MA: Harvard University Press.

Hausch, D. B., Ziemba, W. T., & Rubenstein, M. (1981). Efficiency of the market for racetrack betting. *Management Science, 27,* 1435–1452.

Hayes, J. R. M. (1964). Human data processing limits in decision making. In E. Bennett (Ed.), *Information system science and engineering: Proceedings of the First Congress on the Information System Sciences.* New York: McGraw-Hill.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3,* 275–310.

Hazard, T., & Peterson, C. R. (1973). *Odds versus probabilities for categorical events* (Tech. Rep. 73-2). McLean, VA: Decisions and Designs, Inc.

Heidelberger, M. (1987). Fechner's indeterminism: From freedom to laws of chance. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 117–156). Cambridge, MA: MIT Press.

Heilbroner, R. (1974). *An inquiry into the human prospect.* New York: Norton.

Hendrick, C., & Constantini, A. F. (1970). Number averaging behavior: A primacy effect. *Psychonomic Science, 19,* 121–122.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366–378.

Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the machiguenga of the Peruvian Amazon. *The American Economic Review, 90,* 973–979.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of Homo economicus: Behavioral experiments in k15 small-scale societies. *American Economic Review, 91,* 73–78.

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics, 54,* 791–798.

Herman, L. M., Ornstein, G. N., & Bahrick, H. P. (1964). Operator decision performance using probabilistic displays of object location. *IEEE Transactions on Human Factors in Electronics, 5,* 13–19.

Hertwig, R., & Gigerenzer, G. (1994). Unpublished manuscript. (Described in Gigerenzer, 1994)

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied, 5,* 205–221.

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review, 87,* 398–410.

Hintzman, D. L. (1986). "Schemata abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hintzman, D. L., Asher, S. J., & Stern, L. D. (1978). Incidental retrieval and memory for coincidences. In M. M. Grunberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 61–68). London: Academic Press.

Hirshleifer, D., & Rasmusen, E. (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior and Organization, 12,* 87–106.

Hoch, S. J. (1984). Availability and inference in predictive judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 649–662.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 719–731.

Hoch, S. J., Malcus, L., & Hasher, L. (1986). Frequency discrimination: Assessing global-level and element-level units in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 232–240.

Hoffman, M. (Ed.). (1991). *The world almanac and book of facts.* New York: World Almanac.

Hoffman, P. (1988). *Archimedes revenge: The joys and perils of mathematics.* New York: Fawcett Crest.

Hoffrage, U. (2000). Why the analyses of cognitive processes matter. *Behavioral and Brain Sciences, 23,* 679–680.

Hofstadter, D. R. (1983). Metamagical themas: Computer tournaments of the prisoner's dilemmas suggest how cooperation evolves. *Scientific American, 248*(5), 14–20.

Hofstatter, P. R. (1939). Über die Schatzung von gruppeneigenschaften [On the estimation of group properties]. *Zeitschrift für Psychologie, 145,* 1–44.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association, 70,* 271–291.

Hogarth, R. M. (1987). *Judgement and choice* (2nd ed.). London: Wiley.

Holland, B. K. (1998). *Probability without equations: Concepts for clinicians.* Baltimore, MD: Johns Hopkins University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Hollingworth, H. L. (1913a). Characteristic differences between recall and recognition. *American Journal of Psychology, 24,* 532–544.

Hollingworth, H. L. (1913b). Experimental studies in judgment. *Archives of Psychology, 29,* 1–96.

Hombas, V. C. (1997). Waiting time and expected waiting time—paradoxical situations. *The American Statistician, 51,* 130–133.

Hooke, R. (1989). Basketball, baseball, and the null hypothesis. *Chance: New Directions for Statistics and Computing, 2*(4), 35–37.

Hoorens, V., & Buunk, B. P. (1993). Social comparison of health risks: Locus of control, the person-positivity bias, and unrealistic optimism. *Journal of Applied Social Psychology, 18,* 291–302.

Horgan, J. (1994). Radon's risks. *Scientific American, 271*(2), 14–16.

Horváth, R. A. (1987). The rise of macroeconomic calculations in economic statistics. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 147–169). Cambridge, MA: MIT Press.

Horwich, P. (1982). *Probability and evidence.* New York: Cambridge University Press.

Houghton, R. A., & Woodwell, G. M. (1989). Global climatic change. *Scientific American, 260*(4), 36–44.

Howard, R. A. (1966). Decision analysis: Applied decision theory. In D. B. Hertz & J. Melese (Eds.), *Proceedings of the Fourth International Conference on Operational Research* (pp. 55–71). New York: Wiley-Interscience.

Howard, R. A. (1968). The foundations of decision analysis. *IEEE Transactions on Systems Science and Cybernetics, SSC-4,* 211–219.

Howell, W. C. (1966). Task characteristics in sequential decision behavior. *Journal of Experimental Psychology, 71,* 124–131.

Howell, W. C. (1973). Storage of events and event frequency: A comparison of two paradigms in memory. *Journal of Experimental Psychology, 98,* 260–263.

Howson, C., & Urbach, P. (1993) *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court.

Hoyle, F. (1983). *The intelligent universe.* London: Michael Joseph.

Huberty, C. J., & Morris, J. D. (1988). A single contrast test procedure. *Educational and Psychological Measurement, 48,* 567–578.

Huff, D. (1973). *How to lie with statistics.* New York: Viking Penguin. (Original work published 1954)

Hunter, J. E. (1997). Need: A ban on the significance test. *Psychological Science, 8,* 3–7.

Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97,* 251–273.

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to non-technical people. *Risk Analysis, 7,* 519–529.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books. (Original work published 1955)

Irwin, F. W. (1944). The realism of expectations. *Psychological Review, 51,* 120–126.

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality, 21,* 329–335.

Irwin, F. W., & Snodgrass, J. G. (1966). Effects of independent and dependent outcomes on bets. *Journal of Experimental Psychology, 71,* 282–285.

Jacoby, L. L. (1972). Context effects on frequency judgments of words and sentences. *Journal of Experimental Psychology, 94,* 255–260.

Jacoby, L. L., Bjork, R. A., & Kelley, C. M. (1994). Illusions of comprehension, competence, and remembering. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 57–80). Washington, DC: National Academy Press.

Jaffe, A. J., & Spirer, H. F. (1987). *Misused statistics: Straight talk for twisted numbers.* New York: Marcel Dekker.

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *Journal of Experimental Psychology, 41,* 291–297.

Jaynes, E. P. (1968). Prior probabilities. *IEEE Transactions on System Science and Cybernetics, SSC-4,* 227–240.

Jeffrey, L. (1989). Writing and rewriting poetry: William Wordsworth. In D. B. Wallace & H. E. Gruber (Eds.), *Creative people at work* (pp. 69–89). New York: Oxford University Press.

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: University of Chicago Press.

Jeffreys, H. (1934). Probability and scientific method. *Proceedings of the Royal Society of London, Series A, 146,* 9–15.

Jeffreys, H. (1939). *Theory of probability.* Oxford, England: Clarendon.

Jemmott, J. B., Ditto, P. H., & Croyle, R. T. (1986). Judging health status: Effects of perceived prevalence and personal relevance. *Journal of Personality and Social Psychology, 50,* 899–905.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79* (Whole No. 594).

Jenkins, J. J., & Tuten, J. T. (1992). Why isn't the average child from the average family?—and similar puzzles. *American Journal of Psychology, 105,* 517–526.

Jennings, D., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In A. Tversky, D. Kahneman, & P. Slovic (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). New York: Cambridge University Press.

Jepson, C., Krantz, D. H., & Nisbett, R. E. (1983). Inductive reasoning: Competence or skill? *Behavioral and Brain Sciences, 6,* 94–101.

Johnson, D. (1981). *V-1, V-2: Hitler's vengeance on London.* New York: Stein & Day.

Johnson, E. J., & Schkade, D. A. (1989). Bias in utility assessment: Further evidence and explanations. *Management Science, 35,* 406–424.

Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology, 45,* 20–31.

Johnson, E. M. (1973). Numerical encoding of qualitative expressions of uncertainty (Tech. Paper No. 250). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, E. M. (1974). The effect of data source reliability on intuitive inference (ARI Tech. Paper No. 251, AD 784097). Alexandria, VA: U.S. Army.

Johnson, E. M., Cavanagh, R. C., Spooner, R. L., & Samet, M. G. (1973). Utilization of reliability measurements in Bayesian inference: Models and human performance. *IEEE Transactions on Reliability, R-22,* 176–183.

Johnson, M. K., Peterson, M. A., Yap, E. C., & Rose, P. M. (1989). Frequency judgments: The problem of defining a perceptual event. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 126–136.

Johnson, M. K., Raye, C. L., Wang, A. Y., & Taylor, J. H. (1979). Fact and fantasy: The roles of accuracy and variability in confusing imaginations with perceptual experiences. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 229–240.

Johnson, M. K., Taylor, T. H., & Raye, C. L. (1977). Fact and fantasy: The effects of internally generated events on the apparent frequency of externally generated events. *Memory and Cognition, 5,* 116–122.

Johnston, W. A. (1967). An individual performance and self-evaluation in a simulated team. *Organizational Behavior and Human Performance, 2,* 309–328.

Jonides, J., & Jones, C. M. (1992). Direct coding for frequency of occurrence. *Journal of Experimental Psychology: Leaning, Memory, and Cognition, 18,* 368–378.

Jorland, G. (1987). The St. Petersburg paradox, 1713–1937. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 157–190). Cambridge, MA: MIT Press.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5,* 55–71.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226–246.

Kac, M. (1964). Probability. *Scientific American, 211*(3), 92–108.

Kac, M. (1983). Marginalia: What is random. *American Scientist, 71,* 405–406.

Kahan, J. P. (1974). Rationality, the prisoner's dilemma, and population. *Journal of Social Issues, 30,* 189–210.

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences, 232,* 681–683.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy, 98,* 1325–1348.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39,* 17–30.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972a). On the psychology of prediction. *Oregon Research Institute Bulletin, 12* (Whole No. 4).

Kahneman, D., & Tversky, A. (1972b). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kahneman, D., & Tversky, A. (1979a). On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition, 7,* 409–411.

Kahneman, D., & Tversky, A. (1979b). Prospect theory. An analysis of decision under risk. *Econometrica, 47,* 263–291.

Kahneman, D., & Tversky, A. (1982a). On the study of statistical intuitions. *Cognition, 11,* 123–141.

Kahneman, D., & Tversky, A. (1982b). Variants of uncertainty. *Cognition, 11,* 143–157.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39,* 341–350.

Kalbfleisch, J. G., & Sprott, D. A. (1976). On tests of significance. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2, pp. 259–272). Dordrecht, Netherlands: Reidel.

Kamlah, A. (1987). The decline of the Laplacian theory of probability: A study of Stumpf, von Kries, and Meinong. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 91–116). Cambridge, MA: MIT Press.

Kaplan, M. F., & Schwartz, S. (Eds.). (1977). *Human judgment and decision processes in applied settings.* New York: Academic Press.

Kaplan, R. J., & Newman, J. R. (1966). Studies in probabilistic information processing. *IEEE Transactions in Human Factors in Electronics, 7,* 49–63.

Kareev, Y. (1995a). Positive bias in the perception of covariation. *Psychological Review, 102,* 490–502.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56,* 263–269.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives.* New York: Wiley.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32,* 1–24.

Kellogg, R. T. (1981). Feature frequency in concept learning: What is counted? *Memory and Cognition, 9,* 157–163.

Kempton, W. (1991). Lay perspectives on global climate change. *Global Environmental Change, 1,* 183–208.

Keren, G. B. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39,* 98–114.

Keren, G. B. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Keren, G. B,. & Wagenaar, W. A. (1985). On the psychology of playing blackjack: Normative and descriptive considerations with implications for decision theory. *Journal of Experimental Psychology: General, 114,* 133–158.

Kerr, R. A. (2000). Draft report affirms human influence. *Science, 288,* 589–590.

Kessler, D. A., & Feiden, K. L. (1995). Faster evaluation of vital drugs. *Scientific American, 272*(3), 26–32.

Keynes, J. M. (1921). *Treatise on probability.* London: Macmillan.

Keynes, J. M. (1956). The application of probability to conduct. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 2, pp. 1360–1373). New York: Simon & Schuster. (Original work published 1921)

Kidd, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica, 34,* 338–347.

King, J. P. (1992). *The art of mathematics.* New York: Fawcett Columbine.

Kintsch, W. (1970). *Learning, memory, and conceptual processes.* New York: Wiley.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology, 63,* 534–544.

Klahr, D. (1976). The social psychologist as troll. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 243–250). Hillsdale, NJ: Lawrence Erlbaum Associates.

Klatzky, R. L., & Messick, D. M. (1995). Curtailing medical inspections in the face of negative consequences. *JEP: Applied, 1,* 163–178.

Knight, F. H. (1921). *Risks, uncertainty and profit.* Boston: Houghton Mifflin.

Knobloch, E. (1987). Emile Borel as a probabilist. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 215–233). Cambridge, MA: MIT Press.

Koestler, A. (1972). *The roots of coincidence: An excursion into parapsychology.* New York: Vintage Books.

Kogan, N., & Wallach, M. A. (1967). Risk taking as a function of the situation, the person, and the group. In *New directions in psychology* (Vol. 3, pp. 111–278). New York: Holt, Rinehart & Winston.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1,* 1–7.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

Kramer, G. P., Kerr, N. L., & Carroll, J. S. (1990). Pretrial publicity, judicial remedies, and jury bias. *Law and Human Behavior, 14,* 409–438.

Krantz, D. H. (1981). Improvements in human reasoning and an error in L. J. Cohen's. Commentary on Cohen, 1981. *The Behavioral and Brain Sciences, 4,* 340–341.

Krantz, D. H. (1991). From indices to mappings: The representational approach to measurement. In D. Brown & J. Smith (Eds.), *Frontiers of mathematical psychology* (pp. 1–52). New York: Springer-Verlag.

Kreps, D. M. (1990). *A course in microeconomic theory.* Princeton, NJ: Princeton University Press.

Krischer, J. P. (1980). An annotated bibliography of decision analytic applications to health care. *Operations Research, 28,* 97–113.

Krüger, L. (1987a). The probabilistic revolution in physics: An overview. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 373–378). Cambridge, MA: MIT Press.

Krüger, L. (1987b). The slow rise of probabilism: Philosophical arguments in the nineteenth century. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 60–89). Cambridge, MA: MIT Press.

Krüger, L., Daston, L. J., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Ideas in history* (Vol. 1). Cambridge, MA: MIT Press.

Krüger, L., Gigerenzer, G., & Morgan, M. S. (Eds.). (1987). *The probabilistic revolution: Vol. 2. Ideas in the sciences.* Cambridge, MA: MIT Press.

Kubovy, M., & Gilden, D. L. (1990). Apparent randomness is not always the complement of apparent order. In G. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 115–127). Washington, DC: American Psychological Association.

Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes, 62,* 230–240.

Kuhn, D., Phelps, E., & Walters, J. (1985). Correlational reasoning in an everyday context. *Journal of Applied Developmental Psychology, 6,* 85–97.

Kuhn, D., Weinstock, M., & Flaton, R. (1994). How well do jurors reason? Competence dimensions of individual variations in a juror reasoning task. *Psychological Science, 5,* 289–296.

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology, 53,* 636–647.

Kunda, Z., & Nisbett, R. E. (1986a). Prediction and the partial understanding of the law of large numbers. *Journal of Social Psychology, 22,* 339–354.

Kunda, Z., & Nisbett, R. E. (1986b). The psychometrics of everyday life. *Cognitive Psychology, 18,* 199–224.

Landman, J., & Manis, M. (1983). Social cognition: Some historical and theoretical perspectives. *Advances in Experimental Social Psychology, 16,* 49–123.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32,* 311–328.

Langer, E. J., & Roth, J. (1975). Heads I win, tails is chance: The illusion of control is a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology, 32,* 951–955.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements [Memoir on the probability of causes of events]. *Mé moires pré senté es à l'Acadé mie des Sciences, 6,* 621–656.

Laplace, P. S. (1812). Thé orie analytique des probabilité s [Analytical theory of probabilities]. (Reprinted in Laplace, P. S. [1878–1912]. *Oeuvres complètes de Laplace.* Paris: Gauthier-Villars.)

Laplace, P. S. (1951). *A philosophical essay on probabilities.* (F. W. Truscott & F. L. Emory, Trans.). New York: Dover. (Original work published 1814)

Larkey, P. D., Smith, R. A., & Kadane, J. B. (1989). It's okay to believe in the "hot-hand." *Chance: New Directions for Statistics and Computing, 2*(4), 22–30.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822–832.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology, 73,* 498–502.

Lattimer, J. (1966). *Similarities in fatal woundings of John Wilkes Booth and Lee Harvey Oswald.* New York: s.n.

Lattimer, J. (1980). *Kennedy and Lincoln: Medical and ballistic comparisons of their assassinations.* New York: Harcourt Brace Jovanovich.

Layzer, D. (1990). *Cosmogenesis: The growth of order in the universe.* New York: Oxford University Press.

Lécuyer, B-P. (1987). Probability in vital and social statistics: Quetelet, Farr, and the Bertillons. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 317–335). Cambridge, MA: MIT Press.

Leddo, J., Abelson, R. P., & Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology, 47,* 933–943.

Lederman, L. (1993). *The God particle: If the universe is the answer, what is the question?* New York: Dell.

Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology, 40,* 1435–1441.

Levi, I. (1981). Should Bayesians sometimes neglect base rates? *Behavioral and Brain Sciences, 4,* 342–343.

Levi, I. (1983). Who commits the base rate fallacy? *Behavioral and Brain Sciences, 6,* 502–506.

Levin, B. (1981). A representation for multinomial cumulative distribution functions. *The Annals of Statistics, 9,* 1123–1126.

Levin, I. P. (1974). Averaging processes and intuitive statistical judgments. *Organizational Behavior and Human Performance, 12,* 83–91.

Levin, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General, 104,* 39–53.

Levin, I. P., Chapman, D. P., & Johnson, R. D. (1988). Confidence in judgments based on incomplete information: An investigation using both hypothetical and real gambles. *Journal of Behavioral Decision Making, 1,* 29–41.

Levin, I. P., Johnson, R. D., Deldin, P. J., Carsten, L. M., Cressey, L. J., & Davis, C. D. (1986). Framing effects in decisions with completely and incompletely described alternatives. *Organizational Behavior and Human Decision Processes, 38,* 48–64.

Levin, I. P., Johnson, R. D., Russo, C. P., & Deldin, P. J. (1985). Framing effects in judgment tasks with varying amounts of information. *Organizational Behavior and Human Decision Processes, 36,* 362–377.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 275–324). Dordrecht, Netherlands: Reidel.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101,* 16–20.

Lichtenstein, S., Slovic, P., Fischoff, B. Layman, M., & Coombs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551–578.

Liebrand, W. B. G., Messick, D. M., & Wolters, F. J. M. (1986). Why we are fairer than others: A cross-cultural replication and extension. *Journal of Experimental Social Psychology, 22,* 590–604.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint.* Cambridge, England: Cambridge University Press.

Lindley, D. V. (1971). *Making decisions.* London: Wiley.

Lindley, D. V. (1972). Comment. *Journal of the American Statistical Association, 67,* 373–374.

Lindley, D. V. (1984). A Bayesian lady tasting tea. In H. A. David & H. T. David (Eds.), *Statistics: An appraisal* (pp. 455–485). Ames: Iowa State University Press.

Lindley, D. V. (1993). *The end of physics: The myth of a unified theory.* New York: Basic Books.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics, 9,* 45–58.

Linville, P. W., Fischer, G. W., & Fischhoff, B. (1993). AIDS risk perceptions and decision biases. In J. B. Pryor & G. D. Reeder (Eds.), *The social psychology of HIV infection* (pp. 5–38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1988). The measurement of patients' values in medicine. In J. Dowie & A. Elstein (Eds.), *Professional judgment. A reader in clinical decision making* (pp. 395–408). Cambridge, England: Cambridge University Press. (Original work published 1982)

Lloyd, A. L. (1995). Computing bouts of the prisoner's dilemma. *Scientific American, 272*(6), 110–115.

Loftus, E. F. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Loftus, E. F., & Wagenaar, W. A. (1988). Lawyers' predictions of success. *Jurimetrics Journal, 29,* 437–453.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102–105.

Loftus, G. R. (1995). Data analysis as insight. *Behavior Research Methods, Instruments, and Computers, 27,* 57–59.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Loomes, C., & Sugden, R. (1983). A rationale for preference reversal. *American Economic Review, 73,* 428–432.

Lopes, L. L. (1976). Model-based decision and inference in stud poker. *Journal of Experimental Psychology: General, 105,* 217–239.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 626–636.

Lopes, L. L., & Ekberg, P-H. S. (1980). Test of an ordering hypothesis in risky decision-making. *Acta Psychologica, 45,* 161–167.

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 392–400.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16,* 317–323.

Lowrance, W. (1976). *Of acceptable risk.* San Francisco: Freeman.

Luce, R. (1956). Semi-orders and a theory of utility discrimination. *Econometrica, 24,* 178–191.

Luce, R. D. (2000). Fast, frugal, and surprisingly accurate heuristics. *Behavioral and Brain Sciences, 23,* 757–758.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey.* New York: Wiley.

Lund, A. M., Hall, J. W., Wilson, K. P., & Humphreys, M. S. (1983). Frequency judgment accuracy as a function of age and school achievement (learning disabled versus non-learning-disabled) patterns. *Journal of Experimental Child Psychology, 35,* 236–247.

Lund, F. H. (1925). Psychology of belief. *Journal of Abnormal Social Psychology, 20,* 180–190.

Lunt, P. K., & Livingstone, S. M. (1989). Psychology and statistics: Testing the opposite of the idea you first thought of. *The Psychologist, 2,* 528–531.

Lusted, L. B. (1977). *A study of the efficacy of diagnostic reaiologic procedures: Final report on diagnostic efficacy.* Chicago: Efficacy Study Committee of the American College of Radiology.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Lyon, D., & Slovic, P. (1975). On the tendency to ignore base rates when estimating probabilities. *ORI Research Bulletin, 15,* 1.

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40,* 287–298.

MacCrimmon, K. R. (1968). Decision making among multiple-attribute alternatives: A survey and consolidated approach (RAND Corporation Memo RM 4823-ARPA). Santa Monida, CA: RAND Corporation.

MacDonald, A. B., Klein, J. R., & Wark, D. L. (2003). Solving the solar neutrino problem. *Scientific American, 188*(4), 40–49.

Macdonald, R. R. (1986). Credible conceptions and implausible probabilities. *British Journal of Mathematical and Statistical Psychology, 39,* 15–27.

Macdonald, R. R. (1997a). Base rates and randomness. *Behavioral and Brain Sciences, 20,* 778.

Macdonald, R. R. (1997b). On statistical testing in psychology. *British Journal of Psychology, 88,* 333–347.

Macdonald, R. R., & Gilhooly, K. J. (1990). More about Linda *or* conjunction in context. *European Journal of Cognitive Psychology, 2,* 57–70.

MacGregor, D., Fischhoff, B., & Blackshaw, L. (1987). Search success and expectations with a computer interface. *Information Processing and Management, 23,* 419–432.

Mackie, J. L. (1981). Propensity, evidence, and diagnosis. *Behavioral and Brain Sciences, 4,* 345–346.

Mackworth, N. H. (1965). Originality. *The American Psychologist, 20,* 51–66.

Malakoff, D. (1999). Bayes offers a "new" way to make sense of numbers. *Science, 286,* 1460–1464.

Malgady, R. G. (1998). In praise of value judgments in null hypothesis testing ... and of "accepting" the null hypothesis. *American Psychologist, 53,* 797–798.

Malkiel, B. G. (1985). *A random walk down Wall Street* (4th ed.). New York: Norton.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38,* 231–240.

Marcus, R. (1980). Moral dilemmas and consistency. *Journal of Philosophy, 77*(3), 121–136.

Margalit, A., & Bar-Hillel, M. (1981). The irrational, the unreasonable, and the wrong. Commentary on Cohen, 1981. *The Behavioral and Brain Sciences, 4,* 346–349.

Margolis, H. (1987). *Patterns, thinking, and cognition.* Chicago: University of Chicago Press.

Margolis, H. (2000). Simple heuristics that make us dumb. *Behavioral and Brain Sciences, 23,* 758.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy, 60,* 151–158.

Marks, R. (1951). The effect of probability, desirability, and "privilege" on the stated expectations of change. *Journal of Personality, 19,* 332–351.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 119–140). New York: Oxford University Press.

Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 169–188). New York: Oxford University Press.

Martin, E. (1981). Simpson's paradox resolved: A reply to Hintzman. *Psychological Review, 88,* 372–374.

Martin-Löf, P. (1966). The definition of random sequences. *Information and Control, 9,* 602–619.

Mathematical Sciences Education Board (1988). *A framework for the revision of K–12 mathematical curricula* (Final report to the Mathematical Sciences Education Board from its Curriculum Frameworks Task Force). Washington, DC: National Academy Press.

May, R. S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B Brehmer, H. Jungermann, P. Lourens, & G. Sevon (Eds.), *New directions in research on decision making* (pp. 175–189). Amsterdam: Elsevier Science.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47–60.

McCain, R. A. (2000). Differences, games and pluralism. *Behavioral and Brain Sciences, 23,* 688–689.

McClintock, C. G., & McNeel, S. P. (1966). Reward level and game-playing behavior. *Journal of Conflict Resolution, 10,* 98–102.

McCormick, N. J. (1981). *Reliability and risk analysis.* New York: Academic Press.

McKenna, F. P. (1993). It won't happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology, 84,* 39–50.

McKenna, F. P., Stanier, R. A., & Lewis, C. (1991). Factors underlying illusory self-assessment of driving skill in males and females. *Accident Analysis and Prevention, 23,* 45–52.

McKenzie, C. R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26,* 209–239.

McLeish, J. (1994). *The story of numbers: How mathematics has shaped civilization.* New York: Fawcett Columbine. (Original work published 1991 under the title *Number*)

McNeil, B., Pauker, S., Sox, H., Jr., & Tversky, A. (1982). Comment on the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306,* 1259–1262.

McPhee, J. (1992). *The deltoid pumpkin seed.* New York: Noonday Press. (Original work published 1973)

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Meehl, P. E. (1986). Superiority of actuarial to clinical prediction. *Journal of Personality Assessment, 50,* 370–375.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Molaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391–423). Mahwah, NJ: Lawrence Erlbaum Associates.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52,* 87–106.

Mehle, T., Gettys, C. V., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica, 49,* 127–140.

Ménard, C. (1987). Why was there no probabilistic revolution in economic thought? In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 139–146). Cambridge, MA: MIT Press.

Mendeleev, D. (1956). Periodic law of the chemical elements. In J. R. Newman, (Ed.), *The world of mathematics* (Vol. 1, pp. 913–918). New York: Simon & Schuster. (Original work published circa 1869)

Merton, R. K. (1961). Singletons and multiples in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 105,* 470–486.

Messick, D. M. (1973). To join or not to join: An approach to the unionization decision. *Organizational Behavior and Human Performance, 10,* 145–156.

Messick, D. M., Bloom, S., Boldizar, J. P., & Samuelson, C. D. (1985). Why we are fairer than others. *Journal of Experimental Social Psychology, 21,* 480–500.

Messick, D. M., & Campos, F. T. (1972). Training and conservatism in subjective probability. *Journal of Experimental Psychology, 94,* 335–337.

Messick, D. M., & van de Geer, J. P. (1981). A reversal paradox. *Psychological Bulletin, 90,* 582–593.

Metalsky, G. I., & Abramson, L. Y. (1981). Attributional styles: Toward a framework for conceptualization and assessment. In P. C. Kendall & S. D. Hollon (Eds.), *Assessment strategies for cognitive-behavioral interventions* (pp. 15–58). New York: Academic Press.

Metz, K. H. (1987). Paupers and numbers: The statistical argument for social reform in Britain during the period of industrialization. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 337–350). Cambridge, MA: MIT Press.

Meux, E. P. (1973). Concern for the common good in an N-person game. *Journal of Personality and Social Psychology, 28,* 414–418.

Midden, C. J. H., & Verplanken, B. (1990). The stability of nuclear attitudes after Chernobyl. *Journal of Environmental Psychology, 10,* 111–119.

Miller, D. W., & Starr, M. K. (1967). *The structure of human decisions.* Englewood Cliffs, NJ: Prentice-Hall..

Mintzberg, H. (1975, July/August). The manager's job: Folklore and fact. *Harvard Business Review,* pp. 49–61.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond *deja vu* in the search for cross-situational consistency. *Psychological Review, 89,* 730–755.

Monod, J. (1972). *Chance and necessity.* London: Collins.

Montmort, P. R. (1713). *Essai d'analyse sur les jeux de hasard [Essay on analysis of games of chance].* Paris.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: National Academy Press.

Moore, D. W. (1992). *The super pollsters.* New York: Four Walls Eight Windows.

Morgan, J. P., Chaganty, N. R., Dahiya, R. C., & Doviak, M. J. (1991a). Let's make a deal: The player's dilemma. *The American Statistician, 45,* 284–287.

Morgan, J. P., Chaganty, N. R., Dahiya, R. C., & Doviak, M. J. (1991b). Rejoinder. *The American Statistician, 45,* 289.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.* New York: Cambridge University Press.

Morgan, M. S. (1987). Statistics without probability and Haavelmo's revolution in econometrics. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 171–197). Cambridge, MA: MIT Press.

Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: a task specific phenomenon? *Personality and Social Psychology Bulletin, 10,* 243–252.

Morris, R. (1987). *The nature of reality.* New York: Noonday Press.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago, Aldine.

Moser, P. K. (1990). Rationality in action: General introduction. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 1–10). New York: Cambridge University Press.

Mosteller, F. (1965). *Fifty challenging problems in probability with solutions.* Reading, MA: Addison-Wesley.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Mahwah, NJ: Lawrence Erlbaum Associates.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595–600.

Murphy, A. H. (1974). A sample skill score for probability forecasts. *Monthly Weather Review, 102,* 48–55.

Murphy, A. H., & Katz, R. W. (Eds.) (1983). *Probability, statistics, and decision making in the atmospheric sciences.* Boulder, CO: Westview.

Murphy, A. H., & Winkler, R. L. (1971). Forecasters and probability forecasts: Some current problems. *Bulletin of the American Meteorological Society, 52,* 239–247.

Murphy, A. H., & Winkler, R. L. (1974). Subjective probability forecasting experiments in meteorology: Some preliminary results. *Bulletin of the American Meteorological Society, 55,* 1206–1216.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest, 2,* 2–9.

Murray, D. J. (1987). A perspective for viewing the integration of probability theory into psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 73–100). Cambridge, MA: MIT Press.

Myers, D. G. (2002). *Intuition: The powers and perils of our inner knowing.* New Haven, CT: Yale University Press.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602–627.

Myers, J. C. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 3. Approaches to human learning and motivation* (pp. 171–206). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nagel, E. (1956). The meaning of probability. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 2, pp 1398–1414). New York: Simon & Schuster. (Original work published 1936)

National Academy of Sciences. (1983). *Changing climate: Report of the carbon dioxide assessment committee.* Washington, DC: National Academy Press.

National Aeronautics and Space Administration. (1985). *Space shuttle data for planetary mission RTG safety analysis.* Huntsville, AL: Author.

National Council of Teachers of Mathematics. (1980). *Agenda for action: Recommendations for school mathematics of the 1980s.* Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1989). *Improving risk communication.* Committee on Risk Perception and Communication, National Research Council. Washington, DC: National Academy Press.

Navon, D. (1978). The importance of being conservative: Some reflections on human Bayesian behavior. *British Journal of Mathematical and Statistical Psychology, 31,* 33–48.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Newby-Clark, I. R., Ross, M., Buehler, R, Koehler, D. J., & Griffin, D. (2000). People focus on optimistic scenarios and disregard pessimistic scenarios while predicting task completion times. *Journal of Experimental Psychology: Applied, 6,* 171–182.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika, 20A,* 175–240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika, 20A,* 263–294.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probability *a priori*. *Proceedings of the Cambridge Philosophical Society, 29,* 492–510.

Nickerson, R. S. (1967). Expectancy, waiting time and the psychological refractory period. In A. F. Sanders (Ed.), *Attention and performance* (pp. 23–34). Amsterdam: North-Holland.

Nickerson, R. S. (1994). Electronic bulletin boards: A case study of computer-mediated communication. *Interacting with Computers, 6,* 117–134.

Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin, 120,* 410–433.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125,* 737–759.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Research, 10,* 168–172.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109,* 330–357.

Nickerson, R. S., & Burnham, D. W. (1969). Response times with nonaging foreperiods. *Journal of Experimental Psychology, 79,* 452–457.

Nickerson, R. S., & McGoldrick, C. C. (1963). Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Perceptual and Motor Skills, 17,* 159–167.

Nickerson, R. S., & McGoldrick, C. C. (1965). Confidence ratings and level of performance on a judgmental task. *Perceptual and Motor Skills, 20,* 311–316.

Niiniluoto, I. (1981). L. J. Cohen versus Bayesianism. *Behavioral and Brain Sciences, 4,* 349.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology, 32,* 932–943.

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. (1976). Popular induction: Information is not necessarily informative. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 113–134). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238,* 625–631.

Nisbett, R. E., Krantz, D. H., Jepson, D., & Fong, G. T. (1982). Improving inductive inference. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 445–462). New York: Cambridge University Press.

Nisbett, R. E., Krantz, D. H., Jepson, D., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Schacter, S. (1966). Cognitive manipulation of pain. *Journal of Experimental Social Psychology, 2,* 227–236.

Nogar, R. J. (1966). *The wisdom of evolution.* New York: New American Library.

North, D. W. (1968). A tutorial introduction to decision theory. *IEEE Transactions on Systems Science and Cybernetics, SSC-4,* 200–210.

Novick, M. R., & Jackson, P. E., (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Nowak, M. A., May, R. M., & Sigmund, K. (1995). The arithmetics of mutual help. *Scientific American, 272*(6), 76–79.

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science, 289,* 1773–1775.

Nowak, M. A., & Sigmund, K. (1992). Tit-for-tat in heterogeneous populations. *Nature, 355,* 250–253.

Nozick, R. (1993). *The nature of rationality.* Princeton, NJ: Princeton University Press.

Nuclear Regulatory Commission. (1978). *Risk assessment review group report to the U.S. Nuclear Regulatory Commission* (Rep. No. NUREG/CR-0400). Washington, DC: Author.

Nuclear Regulatory Commission (1986). Safety goals for the operation of nuclear power plants: Policy statement. *Federal Register, 51*(August 4), 28044–28049.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, England: Wiley.

Oberschall, A. (1987). The two empirical roots of social theory and the probability revolution. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 103–131). Cambridge, MA: MIT Press.

O'Connor, M. (1989). Models of human behavior and confidence in judgment: A review. *International Journal of Forecasting, 5,* 159–169.

Office of Technology Assessment, U.S. Congress. (1987). *New developments in biotechnology: Public perceptions of biotechnology.* Washington, DC: U.S. Government Printing Office.

Ogilvy, C. S., & Anderson, J. T. (1966). *Excursions in number theory.* New York: Oxford University Press.

O'Gorman, H. J., & Garry, S. L. (1976). Pluralistic ignorance—a replication and extension. *Public Opinion Quarterly, 40,* 449–458.

Okasha, S. (2000). Bayes, Levi, and the taxicabs. *Behavioral and Brain Sciences, 23,* 693.

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 599–608.

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups.* Cambridge, MA: Harvard University Press.

Ophuls, W. (1973). Leviathan or oblivion? In H. E. Daly (Ed.), *Toward a steady state economy* (pp. 215–230). San Francisco: Freeman.

Osberg, T. M., & Shrauger, J. S. (1986). Self-prediction: Exploring the parameters of accuracy. *Journal of Personality and Social Psychology, 51,* 1044–1057.

Osherson, D. N. (1995). Probability judgment. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (2nd ed., Vol. 3, pp. 35–75). Cambridge, MA: MIT Press.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical production. *Psychological Monographs, 76.*

Oskamp, S. (1965). Overconfidence in case study judgments. *Journal of Consulting Psychology, 29,* 261–265.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review, 86,* 404–417.

Owens, D. (1992). *Causes and coincidences.* Cambridge, England: Cambridge University Press.

Paese, P. W., & Feuer, M. A. (1991). Decisions, actions, and the appropriateness of confidence in knowledge. *Journal of Behavioral Decision Making, 4,* 1–16.

Paese, P. W., & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Organizational Behavior and Human Decision Processes, 48,* 100–130.

Pagels, H. R. (1991). *Perfect symmetry: The search for the beginning of time.* New York: Bantam Books.

Pais, A. (1986). *Inward bound: Of matter and forces in the physical world.* New York: Oxford University Press.

Palmer, T. C., Jr. (2000, February 2). Big Dig costs take a jump of $1.4b. *Boston Globe* (pp. A1, B8).

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman, (Eds.), *Handbook of perception* (Vol. 2, pp. 127–142). New York: Academic Press.

Pareto, V. (1897). *Cours d'économie politique [Course in political economy]* (Vol. 2). Lausanne, Switzerland: Rouge.

Paulos, J. A. (1990). *Innumeracy: Mathematical illiteracy and its consequences.* New York: Vintage Books.

Paulos, J. A. (1992). *Beyond numeracy.* New York: Vintage Books.

Paulos, J. A. (1998). *Once upon a number: The hidden mathematical logic of stories.* New York: Basic Books.

Payne, S. L. (1952). *The art of asking questions.* Princeton, NJ: Princeton University Press.

Pearson, K. & Weldon, W. F. R. (1901). Editorial. *Biometrika, 1,* 3.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Peirce, C. S. (1932). *Collected papers.* (C. Hartshone & P. Weiss, Eds.). Cambridge, MA: Belknap Press.

Peirce, C. S. (1956). The red and the black. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 2, pp. 1334–1340). New York: Simon & Schuster.

Penney, W. (1969). A waiting time problem. *Journal of Recreational Mathematics, 241* (Problems 95 and 96).

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning , Memory, and Cognition, 14,* 521–533.

Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). New York: Cambridge University Press.

Perloff, L. S., & Fetzer, B. K. (1986). Self-other judgments and perceived vulnerability to victimization. *Journal of Personality and Social Psycholgy,50,* 502–510.

Peters, T. (1979, November/December). Leadership: Sad facts and silver linings. *Harvard Business Review,* pp.164–172.

Peterson, C. R. (1980). Recognition of noncontingency. *Journal of Personality and Social Psychology, 38,* 727–734.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulleton, 68,* 29–46.

Peterson, C. R., & DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology, 73,* 61–65.

Peterson, C. R., DuCharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology, 76,* 236–243.

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology, 70,* 117–121.

Peterson, C. R., & Phillips, L. D. (1966). Revision of continuous probability distributions. *IEEE Transactions on Human Factors in Electronics, HFE-7,* 19–22.

Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probability. *Journal of Experimental Psychology, 69,* 522–527.

Peterson, D. R. (1968). *The clinical study of social behavior.* New York: Appleton–Century–Crofts.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37,* 349–360.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41,* 847–855.

Pfleeger, S. L. (1991). *Software engineering: The production of quality software* (2nd ed.). New York: Macmillan.

Phillips, L. D. (1973). *Bayesian statistics for social scientists.* London: Nelson.

Phillips, L. D., & Edwards, W. (1966). Conservation in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–357.

Phillips, L. D., Hays, W. L., & Edwards, W. (1966). Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics, HFE-7,* 7–18.

Phillips, L. D., & Wright, G. N. (1977). Cultural differences in viewing uncertainty and assessing probabilities. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 507–519). Dordrecht, Netherlands: Reidel.

Piattelli-Palmarini, M. (1989). Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition, 31,* 1–44.

Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: Wiley.

Pincus, S. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States, 88,* 2297–2301.

Pincus, S. & Kalman, R. (1997). Not all (possibly) "random" sequences are created equal. *Proceedings of the National Academy of Sciences of the United States, 94,* 3513–3518.

Pincus, S., & Singer, B. H. (1996). Randomness and degrees of irregularity. *Proceedings of the National Academy of Sciences of the United States, 93,* 2083–2088.

Pitz, G. F. (1974). Subjective probability distributions for imperfectly known quantities. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 24–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pitz, G. F., Leung, L. S., Hamilos, C., & Terpening, W. (1976). The use of probabilistic information in making predictions. *Organizational Behavior and Human Performance, 17,* 1–18.

Platt, J. (1973). Social traps. *American Psychologist, 28,* 641–651.

Poincaré, H (1956). Chance. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 2, pp. 1380–1394). New York: Simon & Schuster. (Original work published 1913)

Poisson, S-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées de r>ègles générales du caldul des probabilités [Investigations of the probability of judgments in criminial and civil matters, preceded by general rules of the calculus of probabilities]* Paris: Bachelier.

Polanyi, M. (1963). Experience and the perception of pattern. In K. M. Sayre & F. J. Crossman (Eds.), *The modeling of mind: Computers and intelligence* (pp. 207–220). Notre Dame, IN: Notre Dame Press.

Politser, P. E. (1981). Decision analysis and clinical judgment: A reevaluation. *Medical Decision Making, 1,* 361–389.

Politzer, G., & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research, 20,* 83–103.

Polkinghorne, J. C. (1986). *One world.* London: SPCK.

Polkinghorne, J. (1988). *Science and creation: The searchfor understanding.* Boston: New Science Library Shambhala.

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychological Research and Reviews, 3,* 5–18.

Pollard, P. (1993). How significant is "significance"? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 449–460). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollard, P., & Evans, J. St. B. T. (1983). The effect of experimentally contrived experience on reasoning performance. *Psychological Research, 45,* 287–301.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10,* 159–163.

Pollatsek, A. (1991). Randomness is well-enough understood to be misunderstood. Commentary on "Psychological Conceptions of Randomness." *Journal of Behavioral Decision Making, 4,* 218–220.

Pollatsek, A., Konold, C., Well, A., & Lima, S. (1984). Beliefs underlying random sampling. *Memory and Cognition, 12,* 395–401.

Polya, G. (1954a). *Mathematics and plausible reasoning: Vol. 1. Induction and analogy in mathematics.* Princeton, NJ: Princeton University Press.

Polya, G. (1954b). *Mathematics and plausible reasoning: Vol. 2. Patterns of plausible inference.* Princeton, NJ: Princeton University Press.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Porter, T. M. (1986). *The rise of statistical thinking, 1820–1900.* Princeton, NJ: Princeton University Press.

Porter, T. M. (1987). Lawless society: Social science and the reinterpretation of statistics in Germany, 1850–1880. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 351–375). Cambridge, MA: MIT Press.

Poulton, E. C. (1982). Biases in quantitative judgments. *Applied Ergonomics, 13,* 31–42.

Poundstone, W. (1990). *Labyrinths of reason.* New York: Doubleday.

Poundstone, W. (1992). *Prisoner's dilemma: John von Neumann, game theory, and the puzzle of the bomb.* New York: Anchor Books.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica, 32,* 122–136.

Press, F. (1975). Earthquake prediction. *Scientific American, 232,* 14–23.

Price, D. J. de S. (1961). *Science since Babylon.* New Haven, CT: Yale University Press.

Quetelet, L. A. J. (1829). Recherches statistiques sur le royaume des Pays-Bas [Statistical investigations in the Netherlands]. *Nouveaux mémoires de l'academie royale des sciences et belles-lettres de Bruxelles, 5.*

Quetelet, L. A. J. (1847). De l'influence de libre arbitre de l'homme sur les faits sociaux [On the influence of free will on social acts]. *Bulletin de la Commission Centrale de Statistique* (of Belgium), *3,* 135–155.

Quetelet, L. A. J. (1848). Du systéme social et des lois qui le régissent [On the social system and the laws that govern it]. Paris: Guillaumin.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty.* Reading, MA: Addison-Wesley.

Raiffa, H., & Schlaifer, R. O. (1961). *Applied statistical decision theory.* Cambridge, MA: Harvard Business School.

Rapoport, Am. (1967). Optimal policies for the prisoner's dilemma. *Psychological Review, 74,* 136–145.

Rapoport, Am., & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General, 121,* 352–363.

Rapoport, Am., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology, 23,* 131–176.

Rapoport, An., & Chammah, A. M. (1965). *Prisoner's dilemma.* Ann Arbor: University of Michigan Press.

Rathje, W. L. (1989, December). Rubbish! *The Atlantic Monthly,* pp. 99–109.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). Mahwah, NJ: Lawrence Erlbaum Associates.

Renn, O. (1990). Public response to the Chernobyl accident. *Journal of Environmental Psychology, 10,* 151–167.

Richardson, L. F. (1956). Statistics of deadly quarrels. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 2, pp. 1254–1263). New York: Simon & Schuster.

Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). Mahwah, NJ: Lawrence Erlbaum Associates.

Roberts, L. (1989). Global warming: Blaming the sun. *Science, 246,* 992.

Robertson, L. S. (1977). Car crashes: Perceived vulnerability and willingness to pay for crash protection. *Journal of Consumer Health, 3,* 136–141.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21–26.

Robinson, L. B., & Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 443–456.

Roby, T. B. (1965). Belief states and the uses of evidence. *Behavioral Science, 10,* 255–270.

Ronis, D. L., & Yates, F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40,* 193–218.

Rose, L. E. (1972). Countering a counter-intuitive probability. *Philosophy of Science, 39,* 523–524.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37,* 322–336.

Roth, A. E. (1995). Bargaining experiments. In J. H. Kagel & A. E. Roth (Eds.), *Handbook of experimental economics* (pp. 253–348). Princeton, NJ: Princeton University Press.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review, 81,* 1068–1095.

Rothman, S., & Lichter, S. R. (1996). Is environmental cancer a political disease? In P. R. Gross, N. Levitt, & M. W. Lewis (Eds.), *The flight from reason* (pp. 231–245). New York: New York Academy of Sciences.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin, 119,* 149–158.

Rowen, J. (1973). Effects of some probability displays on choice. *Organizational Behavior and Human Performance, 9,* 1–15.

Roxburgh, I. W. (1977). The cosmical mystery: The relationship between microphysics and cosmology. In R. Duncan & M. Weston-Smith. (Eds.), *The encyclopedia of ignorance* (pp. 37–45). New York: Pocket Books, Simon & Schuster.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Mahwah, NJ: Lawrence Erlbaum Associates.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Ruelle, D. (1991). *Chance and chaos.* Princeton, NJ: Princeton University Press.

Russel, G. D. (1973). *Lincoln and Kennedy: Looked at kindly together.* New York: Carlton Press.

Ryback, D. (1967). Confidence and accuracy as a function of experience in judgment-making in the absence of systematic feedback. *Perceptual and Motor Skills, 24,* 331–334.

Saarinen, T. F. (1980). Reconnaissance trip to Mt. St. Helens, May 18–21, 1980. *The Bridge* (National Academy of Engineering), *10,* 19–22.

Sahlin, N-E. (1987). The significance of empirical evidence for developments in the foundations of decision theory. In D. Batens & J. P. Van Bendegem (Eds.), *Theory and experiment* (pp. 103–122). Dordrecht, Netherlands: Reidel.

Saks, M. J., & Kidd, R. F. (1980). Human information processing and adjudication: Trial by heuristics. *Law and Society Review, 15,* 123–160.

Salmon, W. C. (1974). The pragmatic justification of induction. In R. Swinburne (Ed.), *The justification of induction* (pp. 85–97). London: Oxford University Press.

Salsburg, D. (2001). *The lady tasting tea.* New York: Freeman.

Samet, M. (1975a). Quantitative interpretation of two qualitative scales used to rate military intelligence. *Human Factors, 17,* 76–86.

Samet, M. (1975b). Subjective interpretation of reliability and accuracy scales for evaluating military intelligence (Tech. Paper No. 260). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Samuelson, P. A. (1960). The St. Petersburg paradox as a divergent double limit. *International Economic Review, 1,* 31–37.

Samuelson, P. A. (1977). St. Petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature, 15,* 24–55.

Savage, L. J. (1954). *Foundation of statistics.* New York: Wiley.

Savage, L. J. (1962). Subjective probability and statistical practice. In G. A. Barnard & D. R. Cox (Eds.), *The foundations of statistical inference: A discussion* (pp. 9–35). New York: Wiley.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover. (Original work published 1954)

Savage, P. S., Jr. (1994). The paradox of nontransitive dice. *The American Mathematical Monthly, 101,* 429–436.

Schaffner, P. E. (1985). Specious learning about reward and punishment. *Journal of Personality and Social Psychology, 48,* 1377–1386.

Schkade, D. A., & Johnson, E. J. (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes, 44,* 203–231.

Schlotterbek, M. (1992). *Bayesian inference without base rates.* Diploma thesis, University of Tübingen, Germany. (Described in Gigerenzer, 1994)

Schmidt, F. L. (1992). What do data really mean? *American Psychologist, 47,* 1173–1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmitt, N., Coyle, B. W., & King, L. (1976). Feedback and task predictability as determinants of performance in multiple cue probability learning task. *Organizational Behavior and Human Decision Processes, 16,* 388–402.

Schum, D. A., & DuCharme, W. M. (1971). Comments on the relationship between the impact and the reliability of evidence. *Organizational Behavior and Human Performance, 6,* 111–131.

Schum, D. A., DuCharme, W. M., & DePitts, K. E. (1973). Research on human multi-stage probabilistic inference processes. *Organizational Behavior and Human Performance, 10,* 318–348.

Schum, D. A., Goldstein, I. L., & Southard, J. F. (1966). Research on a simulated Bayesian information processing system. *IEEE Transactions on Human Factors in Electronics, HFE-7,* 37–48.

Schum, D. A., & Martin, A. W. (1980). *Probabilistic opinion revision on the basis of evidence at trial: A Baconian or Pascalian process* (Research Rep. No. 80-02). Houston, TX: Rice University, Department of Psychology.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17,* 105–151.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110,* 101–120.

Schwarz, N. (1995). Social cognition: Information accessibility and use in social judgment. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (2nd ed., Vol. 3, pp. 345–376). Cambridge, MA: MIT Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. I. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 59–82). New York: Seminar Press.

Seaver, D. A., von Winterfeldt, D. V., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance, 21,* 379–391.

Sedikides, C. (1992). Changes in the valence of the self as a function of mood. In M. S. Clark (Ed.), *Emotion and social abehavior: Review of personality and social psychology* (Vol 14., pp. 271–311). Newbury Park, CA: Sage.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Seitz, F., Jastrow, R., & Nierenberg, W. A. (1989). *Scientific perspectives on the greenhouse problem.* Washington, DC.: George C. Marshall Institute.

Selvin, S. (1975a). On the Monty Hall problem. *The American Statistician, 29,* 134.

Selvin, S. (1975b). A problem in probability [Letter to the editor]. *The American Statistician, 29,* 67.

Sennott, C. M., & Palmer, T. C., Jr. (1994, September 11). The big dig. *Boston Globe,* pp. 1, 40, 41, 43.

Seymann, R. G. (1991). Comment. *The American Statistician, 45,* 287–288.

Shackle, G. L. S. (1967). *Time in economics.* Amsterdam: North Holland. (Original work published 1958)

Shafer, G. (1986). Savage revisited (with discussion), *Statistical Science, 1,* 463–501.

Shafer, G. (1993). Can the various meanings of probability be reconciled? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 165–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shafer, G., & Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Science, 9,* 309–339.

Shafir, E. (1993). Intuitions about rationality and cognition. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 260–283). London: Routledge.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24,* 449–474.

Shafir, E., & Tversky, A. (1995). In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (2nd ed., Vol. 3, pp. 77–100). Cambridge, MA: MIT Press.

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory and Cognition, 10,* 520–530.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8,* 208–224.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory and Cognition, 8,* 459–467.

Shapiro, B. J. (1983). *Probability and certainty in seventeenth-century England.* Princeton, NJ: Princeton University Press.

Shapiro, S. H. (1982). Collapsing contingency tables—a geometric approach. *The American Statistician, 36,* 43–46.

Sharlin, H. I. (1986). EDB: A case study in communicating risk. *Risk Analysis, 6,* 61–68.

Shaughnessy, J. M., & Dick, T. (1991). Monty's dilemma: Should you stick or switch? *Mathematics Teacher, 84*(4), 252–256.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61,* 293–316.

Shaw, G. B. (1956). The vice of gambling and the virtue of insurance. In J. R. Newman (Ed.), *The world of mathematics* (pp. 1524–1531). New York: Simon & Schuster. (Original work published 1944)

Shedler, T. & Manis, M. (1986). Can the availability heuristic explain vividness effects? *Journal of Personality and Social Psychology, 51,* 26–36.

Shepard, R. N. (1964). On subjectively optimal selection among multi-attribute alternatives. In M. W. Shelley, II, & G. L. Bryan (Eds.), *Human judgment and optimality* (pp. 257–281). New York: Wiley.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology, 39,* 211–221.

Sherman, S. J., McMullen, M. N., & Gavanski, I. (1992). Natural sample spaces and the inversion of conditional judgments. *Journal of Experimental Social Psychology, 28,* 401–421.

Shodell, M. (1985). Risky business. *Science, 230,* 43–47

Showers, C., & Cantor, N. (1985). Social cognition. A look at motivated strategies. *Annual Review of Psychology, 36,* 275–305.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2.

Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika, 31,* 125–147.

Schweder, R. A. (1977). Likeness and likelihood in everyday thought: Magical thinking in everyday judgments about personality. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 446–467). New York: Cambridge University Press.

Shweder, R. A. & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder (Ed.), *New directions for methodology of behavioral science: Fallible judgment in behavioral research* (pp. 37–58). San Francisco: Jossey Bass.

Sigmund, K., Fehr, E., & Nowak, M. A. (2002). The economics of fair play. *Scientific American, 286*(1), 82–87.

Silka, L., & Albright, L. (1983) Intuitive judgments of rate of change: The use of teenage pregnancies. *Basic and Applied Social Psychology, 4,* 337–352.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69,* 99–118.

Simon, H. A. (1957). *Models of man: Social and rational.* New York: Wiley.

Simon, H. A. (1990). Alternative visions of rationality. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 189–204). New York: Cambridge University Press. (Originally appeared in *Reason in human affairs,* 1983 (pp. 7–35). Stanford, CA: Stanford University Press)

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13,* 238–241.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis confirmatory strategies and perceived hypothesis confirmation. *Journal of Experimental Social Psychology,* 22, 93–121.

Slote, M. (1985). *Commonsense morality and consequentialism.* London: Routledge & Kegan Paul.

Slote, M. (1986). Moderation, rationality, and virtue. In *The Tanner Lectures on Human Values* (Vol. 7). Salt Lake City, UT: University of Utah Press.

Slote, M. (1990). Rational dilemmas and rational supererogation. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches* (pp. 465–480). New York: Cambridge University Press. (Original work published 1986)

Slovic, P. (1966). Value as a determiner of subjective probability. *IEEE Transactions on Human Factors in Electronics, HFE-7,* 22–28.

Slovic, P. (1972). Information processing, situation specificity, and the generality of risk-taking behavior. *Journal of Personality and Social Psychology, 22,* 128–134.

Slovic, P. (1987). Perception of risk. *Science, 236,* 280–285.

Slovic, P. (1995). The construction of preference. *American Psychologist, 50,* 364–371.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 165–184). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28,* 1–39.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1978). Accident probabilities and seat belt usage: A psychological perspective. *Accident Analysis and Prevention, 10,* 281–285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1979). Rating the risks. *Environment, 21*(3), 14–20, 36–39.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1981). Perception and acceptability of risk from energy systems. In A. Baum & J. E. Singer (Eds.), *Advances in environmental psychology, Vol. 3: Energy conservation: Psychological perspectives* (pp. 155–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1986). Informing the public about the risks from ionizing radiation. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision-making: An interdisciplinary reader* (pp. 114–121). New York: Cambridge University Press. (Original work published 1981)

Slovic, P., Flynn, J. H., & Layman, M. (1991). Perceived risk, trust, and the politics of nuclear waste. *Science, 254,* 1603–1607.

Slovic, P., Kunreuther, H. & White, G. F., (1974). Decision processes, rationality, and adjustment to natural hazards. In G. F. White (Ed.), *Natural hazards: Local, national, global* (pp. 187–205). New York: Oxford University Press.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–744.

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *American Economic Review, 73,* 596–605.

Slovic, P., & MacPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance, 11,* 172–194.

Slugoski, B. R., & Wilson, A. E. (1998). Contribution of conversational skills to the production of judgmental errors. *European Journal of Social Psychology, 28,* 575–601.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4,* 165–173.

Smith, D. S. (1979). Averages for units and averages for individuals within units: A note. *Journal of Family History, 4,* 84–86.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, E. E., Osherson, D. A., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485–527.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32,* 25–38.

Smyth, C. P. (1865). *Our inheritance in the great pyramid* (The last of five editions, greatly revised, was published in 1891). New York: Anson D. Randolph & Co.

Snapper, K. J., & Fryback, D. G. (1971). Inference based on unreliable reports. *Journal of Experimental Psychology, 87,* 401–404.

Snapper, K. J., & Peterson, C. R. (1971). Information seeking and data diagnosticity. *Journal of Experimental Psychology, 87,* 429–433.

Sniezek, R. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46,* 264–282.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory and Psychology, 8,* 291–311.

Southard, J. F., Schum, D. A., & Briggs, G. E. (1964). *Subject control over a Bayesian Hypothesis-selection aid in a complex information processing system* (AMRL-TR-64-95). Columbus: Ohio State University

Sowby, F. D. (1965). Radiation and other risks. *Health Physics, 11,* 879–887.

Sprent, P. (1988). *Taking risks: The science of uncertainty.* New York: Viking Penguin.

Sproul, R. C. (1994). *Not a chance: The myth of chance in modern science and cosmology.* Grand Rapids, MI: Baker Books.

Starr, C. (1972). Benefit-cost studies in sociotechnical systems. In National Academy of Engineering, *Perspectives on benefit-risk decision making.* Washington, DC: National Academy Press.

Steadman, H. J., & Cocozza, J. J. (1974). *Careers of the criminally insane: Excessive social control of deviance.* Lexington, MA: Lexington Books.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Molaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–256). Mahwah, NJ: Lawrence Erlbaum Associates.

Steiger, J. H., & Gettys, C. F. (1972). Best-guess errors in multi-stage inference. *Journal of Experimental Psychology, 92,* 1–7.

Steinhaus, H., & Trybula, S. (1959). On a paradox in applied probabilities. *Bulletin, Academie Polonaise de Sciences, 7,* 67–69.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Sternberg, R. J. (1981). Intelligence as thinking and learning skills. *Educational Leadership, 39,* 28–30.

Stevens, P. K. (1974). *Patterns in nature.* Boston, MA: Little, Brown.

Stewart, I. (1987). *The problems of mathematics.* New York: Oxford University Press.

Stewart, I. (1989). Chaos: Does God play dice? *1990 yearbook of science and the future* (pp. 57–73). Chicago: Encyclopedia Britannica.

Stewart, I. (1997). The lore and lure of dice. *Scientific American, 277*(5), 110–113.

Stigler, S. M. (1987). The measurement of uncertainty in nineteenth-century social science. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 287–292). Cambridge, MA: MIT Press.

Stove, D. C. (1986). *The rationality of induction.* Oxford, England: Clarendon.

Strecker, A. (1859). *Theorien und Experimente zur Bestimmung der Atomgewichte der Elemente [Theory and experiment on the atomic weight of elements].* Braunschweig, Germany.

Sundstrom, E., Lounsbury, J. W., DeVault, R. C., & Peele, E. (1981). Acceptance of a nuclear power plant: Applications of the expectancy-value model. In A. Baum & J. E. Singer (Eds.), *Advances in environmental psychology: Vol. 3. Energy conservation: Psychological perspectives* (pp. 171–189). Hillsdale, NJ: Lawrence Erlbaum Associates.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47,* 143–148.

Svenson, O., Fischhoff, B., & MacGregor, D. (1985). Perceived driving safety and seatbelt usage. *Accident Analysis and Prevention, 17,* 119–133.

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers.* New York: Wiley.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000a). Better decisions through science. *Scientific American, 283*(4), 82–87.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340.

Székely, G. J. (1986). *Paradoxes in probability theory and mathematical statistics.* Dordrecht, Netherlands: Reidel.

Tanner, W. P. Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409.

Taylor, J. (1859). *The great pyramid: Why was it built? And who built it?* London: Longman, Green, Lonman & Roberts.

Taylor, S. E. (1982). The availability bias in social perception and interaction. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 190–200). Cambridge, England: Cambridge University Press.

Teigen, K. H. (1988). When are low-probability events judged to be "probable"? Effects of outcome-set characteristics on verbal probability judgments. *Acta Psychologica, 67,* 27–38.

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80,* 155–190.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52,* 700–709.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization, 1,* 39–60.

Thaler, R. (1986). Illusions and mirages in public policy. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision-making: An interdisciplinary reader* (pp. 161–172). New York: Cambridge University Press. (Original work published 1983)

Thomas, L. (1979). *The medusa and the snail: More notes of a biology watcher.* New York: Viking.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29–32.

Thompson, C. P., Cowan, T., Frieman, J., Mahadevan, R. S., Vogl, R. J., & Frieman, J. (1991). Rajan: A study of a mimorist. *Journal of Memory and Language, 30,* 702–724.

Tippett, L. C. (1956). Sampling and standard error. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 3, pp. 1459–1486). New York: Simon & Schuster. (Original work published 1941)

Toda, M. (1963). *Measurement of subjective probability distributions* (Decision Sciences Laboratory Tech. Documentary Rep. No. ESD-TDR-63-407). Bedford, MA: U.S. Air Force.

Todd, P. M., & Miller, G. F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 287–308). New York: Oxford University Press.

Todhunter, I. (2001). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace.* Briston, England: Thoemmes Press. (Original work published 1865)

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, England: Cambridge University Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46,* 35–57.

Trope, Y., & Bassock, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology, 43,* 22–34.

Trope, Y., & Bassock, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology, 19,* 560–576.

Trow, W. C. (1923). Psychology of confidence. *Archives of Psychology, 67,* 7–37.

Tsal, Y. (1977). Symmetry and transitivity assumptions about a non specified logical relation. *Quarterly Journal of Experimental Psychology, 29,* 677–684.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin, 76,* 105–110.

Turner, J. R. (1987). Random genetic drift, R. A. Fisher, and the Oxford school of ecological genetics. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 313–354). Cambridge, MA: MIT Press.

Turner, T. R. (1993). *The Lincoln and Kennedy assassinations in historical context.* Lincoln Memorial Library. Lincoln, ME.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76,* 31–48.

Tversky, A., & Gilovich, T. (1989a). The cold facts about the "hot hand" in basketball. *Chance: New Directions for Statistics and Computing, 2*(1), 16–21.

Tversky, A., & Gilovich, T. (1989b). The "hot hand": Statistical reality or cognitive illusion. *Chance: New Directions for Statistics and Computing, 2*(4), 31–34.

Tversky, A., & Kahneman, D. (1971). Belief in the "Law of small numbers." *Psychological Bulletin, 76,* 105–110.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Tversky, A., & Kahneman, D. (1978). Causal schemata in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review, 80,* 204–217.

Uspensky, J. V. (1937). *Introduction to mathematical probability.* New York: McGraw-Hill.

Vallone, R., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology, 58,* 582–592.

van der Pligt, J., & Midden, C. J. H. (1990). Chernobyl four years later: Attitudes, risk management and communication. *Journal of Environmental Psychology, 10,* 91–99.

van Wallendael, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior and Human Decision Processes, 43,* 385–405.

van Wallendael, L. R., & Hastie, R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory and Cognition, 18,* 240–250.

Vaughn, E., & Nordenstam, B. (1991). The perception of environmental risks among ethnically diverse groups. *Journal of Cross-Cultural Psychology, 22,* 29–60.

Venn, J. (1888). *The logic of chance* (3rd ed.). London: Macmillan.

Verplanken, B. (1989). Beliefs, attitudes, and intentions toward nuclear energy before and after Chernobyl in a longitudinal within-subjects design. *Environment and Behavior, 21,* 371–392.

Vlek, C. (1970). Multiple probability learning: Associating events with their probabilities of occurrence. *Acta Psychologica, 33,* 207–232.

Vlek, C., & Stallen, P. (1980). Rational and personal aspects of risk. *Acta Psychologica, 45,* 273–300.

von Däniken, E. (1969). *Chariots of the gods.* New York: Bantam.

von Mises, R. (1939). Über Aufteilungs—und Besetzungs-Wahrscheinlichkeiten [On partitions—and filling probabilities]. *Revue de la Faculté des Sciences de l'Université d'Istanbul, N. S., 4,* 145–163.

von Mises, R. (1957). *Probability, statistics and truth.* London: Allen & Unwin. (Original work published 1928)

von Neumann, J., & Morgenstern, D. (1953). *The theory of games and economic behavior* (3rd. ed.). New York: Wiley. (Original work published 1944)

von Plato, J. (1987). Probabilistic physics the classical way. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 379–407). Cambridge, MA: MIT Press.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* Cambridge, England: Cambridge University Press.

Vos Savant, M. (1990a, September 9). Ask Marilyn. *Parade Magazine,* p. 15.

Vos Savant, M. (1990b, December 2). Ask Marilyn. *Parade Magazine,* p. 25.

Vos Savant, M. (1991, February 17). Ask Marilyn. *Parade Magazine,* p. 12.

Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological tests: A survey of clinical psychologists. *American Psychologist, 32,* 874–882.

Wagenaar, W. A. (1970). Subjective randomness and the capacity to generate information. In A. F. Sanders (Ed.), *Attention and Performance III, Acta Psychologica, 33,* 233–242.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of the literature. *Psychological Bulletin, 77,* 65–72.

Wagenaar, W. A. (1991). Randomness and randomizers: Maybe the problem is not so big. Commentary on "Psychological Conceptions of Randomness." *Journal of Behavioral Decision Making, 4,* 220–222.

Wagenaar, W. A., & Keren, G. B. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In E. Hollnagel, G. Maneine, & D. Woods (Eds), *Intelligent decision support in process environments* (pp. 87–107). Berlin: Springer.

Wagenaar, W. A., & Keren, G. B. (1988). Chance and luck are not the same. *Journal of Behavioral Decision Making, 1,* 65–75.

Wagenaar, W. A., Keren, G. B., & Pleit-Kuiper, A. (1984). The multiple objectives of gamblers. *Acta Psychologica, 56,* 167–178.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36,* 46–48.

Waldrop, R. L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician, 49,* 24–28.

Walker, L. Thibaut, J., & Andreoli, V. (1972). Order of presentation at trial. *Yale Law Journal, 82,* 216–226.

Wallach, M. A., & Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology, 1,* 1–19.

Waller, R. W., & Keeley, S. M. (1978). Effects of explanation and information feedback on the illusory correlation phenomenon. *Journal of Consulting and Clinical Psychology, 46,* 342–343.

Wallsten, T. S., & Budescu, D. V. (1980). *Encoding subjective probabilities: A psychological and psychometric review.* Report to the Strategies and Standards Division, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science, 29,* 152–173.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115,* 348–365.

Wandersman, A. H., & Hallman, W. K. (1993). Are people acting irrationally? Understand public concerns about environmental threats. *American Psychologist, 48,* 681–686.

Ward, K. (1996). *God, chance and necessity.* Oxford, England: Oneworld.

Ward, W. D., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231–241.

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 509–521.

Watson, S. R., & Buede, D. M. (1987). *Decision synthesis: The principles and practice of decision analysis.* New York: Cambridge University Press.

Weart, S. (1988). *Nuclear fear: A history of images.* Cambridge, MA: Harvard University Press.

Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory and Cognition, 23,* 12–22.

Weaver, C. A., III, Bryant, D. S., & Burns, K. E. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver, III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177–193). Hillsdale, NJ: Lawrence Erlbaum Associates.

Weaver, W. (1950). Probability. *Scientific American, 183*(4), 44–47.

Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, 5,* 49–55.

Weinstein, N. D. (1979). Seeking reassuring or threatening information about environmental cancer. *Journal of Behavioral Medicine, 16,* 220–224.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820.

Weinstein, N. D. (1982). Unrealistic optimism about susceptibility to health problems. *Journal of Behavioral Medicine, 5,* 441–460.

Weinstein, N. D. (1983). Reducing unrealistic optimism about illness susceptibility. *Health Psychology, 2,* 11–20.

Weinstein, N. D. (1984). Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology, 3,* 431–457.

Weinstein, N. D. (1987). Unrealistic optimism about illness susceptibility: Conclusions from a community-wide sample. *Journal of Behavior Medicine, 10,* 481–500.

Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science, 246,* 1232–1233.

Weinstein, N. D., Klotz, M. L., & Sandman, P. (1988). Optimistic biases in public perception of the risk form radon. *American Journal of Public Health, 78,* 796–800.

Weinstein, N. D., & Lachendro, E. (1982). Ego-centrism and unrealistic optimism about the future. *Personality and Social Psychology Bulletin, 8,* 195–200.

Weinstein, N. D., Sandman, P. M., & Roberts, N. E. (1991). Perceived susceptibility and self-protective behavior: A field experiment to encourage home radon testing. *Health Psychology, 10,* 25–33.

Weintraub, R. (1988). A paradox of confirmation. *Erkenntnis, 29,* 169–180.

Weld, H. P., & Danzig, E. R. (1940). A study of the way a verdict is reached by a jury. *American Journal of Psychology, 53,* 518–536.

Weld, H. P., & Roff, M. (1938). A study of the formation of opinion based upon legal evidence. *American Journal of Psychology, 51,* 609–628.

Wells, G. L., & Harvey, J. H. (1977). Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology, 35,* 279–293.

Wheeler, J. A. (1986). Herman Weyl and the unity of knowledge. *American Scientist, 74,* 366–375.

Wheeler, M. (1976). *Lies, damn lies, and statistics: The manipulation of public opinion in America.* New York: Dell.

White, R. M. (1990). The great climate debate. *Scientific American, 263*(1), 36–43.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychological Review, 1,* 3–18.

Wigner, E. P. (1980). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics, 13,* 1–14. (Original work published 1960)

Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences, 72,* 143–146.

Wilson, R. (1979). Analyzing the daily risks of life. *Technology Review, 81,* 40–46.

Wilson, R., & Crouch, E. A. C. (1987). Risk assessment and comparisons: An introduction. *Science, 236,* 267–270.

Wilson, W., Miller, H. L., & Lower, J. S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 68,* 188–196.

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist, 4,* 140–143.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2,* 343–364.

Winefield, A. H. (1966). Negative recency and event dependence. *Quarterly Journal of Experimental Psychology, 18,* 47–54.

Winkler, R. L. (1972). Comment. *Journal of the American Statistical Association, 67,* 376–378.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology, 1,* 751–758.

Winkler, R. L., & Murphy, A. H. (1973). Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance, 10,* 252–270.

Wittgenstein, L. (1972). *On certainty* (G. E. M. Anscombe & G. H. von Wright, Eds.; D. Paul & G. E. M. Anscombe, Trans.). New York: Harper Torchbooks. (Original work published 1953)

Wright, G. N., & Phillips, L. D. (1986). Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 417–431). New York: Cambridge University Press. (Original work published 1980 in the *International Journal of Psychology, 15,* 239–257)

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K .O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology, 9,* 285–299.

Wright, J. C. (1962). Consistency and complexity of response sequences as a function of schedules of noncontingent reward. *Journal of Experimental Psychology, 63,* 601–609.

Wright, W. F., & Bower, G. H. (1992). Mood effects on subjective probability assessment. *Organizational Behavior and Human Decision Processes, 52,* 276–291.

Wrone, D. R. (1981). *Two assassinations: Abraham Lincoln and John F. Kennedy.* Madison: Lincoln Fellowship of Wisconsin.

Wyer, R. S. (1974). *Cognitive organization and change: An information approach.* Potomac, MD: Lawrence Erlbaum Associates.

Wyer, R. S. (1977). The role of logical and nonlogical factors in making inferences about category membership. *Journal of Experimental Social Psychology, 13,* 577–595.

Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132–156.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice-Hall.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–411). Chichester, England: Wiley.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including "signed summation." *Organizational Behavior and Human Decision Processes, 37,* 230–253.

Zabell, S. L. (1981). Unphilosophical probability. *Behavioral and Brain Sciences, 4,* 358–359.

Zabell, S. L. (1988). Symmetry and its discontent. In B. Skyrms & W. L. Harper (Eds.), *Causation, chance and credence* (pp. 155–190). Dordrecht, Netherlands: Kluwer Academic.

Zabell, S. L. (1993). A mathematician comments on models of juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 263–269). New York: Cambridge University Press.

Zacks, R. T., Hasher, L., & Hoch, H. S. (1986). Inevitability and automaticity: A response to Fisk. *American Psychologist, 41,* 216–218.

Zacks, R. T., Hasher, L., & Sanft, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 106–116.

Zakay, D. (1983). The relationship between the probability assessor and the outcome of an event as a determiner of subjective probability. *Acta Psychologica, 53,* 271–280.

Zakay, D. (1984). The influence of perceived event's controllability on its subjective occurrence probability. *The Psychological Record, 34,* 233–240.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley.

Ziskin, J. (1975). *Coping with psychiatric and psychological testimony.* Beverly Hills, CA: Law and Psychology Press.

Author Index

This page intentionally left blank

Subject Index