What Is Probability?[1]

Glenn Shafer

1 Introduction

What is probability? What does it mean to say that the probability of an event is 75%? Is this the frequency with which the event happens? Is it the degree to which we should believe it will happen or has happened? Is it the degree to which some particular person believes it will happen? These questions have been debated for several hundred years. Anyone who teaches statistics should have some sense of how this debate has gone and some respect for the different viewpoints that have been expressed. Each seems to have its germ of truth.

This chapter introduces the debate to those who are not familiar with it. It also sketches a way of reconciling the different viewpoints and draws some lessons for the teacher of probability and statistics.

It is conventional to say that mathematical probability theory has a number of different interpretations. The same mathematical rules (Kolmogorov's axioms and definitions) are obeyed by degrees of belief, by frequencies, and by degrees of evidential support. We can study these rules for their own sake (this is pure mathematics), or we can adopt one of the interpretations and put the rules to use (this is statistics or applied probability). Section 2 explores this conventional formulation. It reviews Kolmogorov's axiomatization and the three standard interpretations of this axiomatization: the belief interpretation, the frequency interpretation, and the support interpretation. Each interpretation, as we shall see, has its appeal and its difficulties.

Section 3 reviews very briefly how the three standard interpretations handle statistical inference. This reveals that they are not as distinct as they first appear. The belief and support interpretations both use Bayesian inference. Moreover, they use Bayesian inference to find beliefs (or degrees of support) about frequentist probabilities. Proponents of the frequency interpretation

---

[1] To appear in *Perspectives on Contemporary Statistics*, edited by David C. Hoaglin and David S. Moore. Mathematical Association of America.

reject Bayesian inference in most cases, but they too end up interpreting certain probabilities as beliefs about frequentist probabilities.

Perhaps frequency, degree of belief, and degree of support are *not* merely three distinct interpretations of the same set of axioms, unrelated except for the coincidence that they follow the same mathematical rules. Perhaps they are more entangled than this—so entangled that it is more accurate to say that they are aspects of a single complex idea. This is the thesis of Section 4, which argues that probability is a complex idea, one that draws together ideas about fair price, rational belief, and knowledge of the long run.

Section 5, in conclusion, draws some lessons for teaching. The most important lesson is humility. Whenever we tell students, “This is what probability really means,” we are wrong. Probability means many things.

2 Three Interpretations of Kolmogorov's Axioms

For the pure mathematician of probability, the axioms and definitions that A.N. Kolmogorov published in 1933 are inseparable from his demonstration that they could be used as a rigorous basis for the study of infinite sequences of random variables. Here, however, we are not interested in infinity. We are interested instead in the implications of Kolmogorov's axiomatization for the meaning of probability, and for this purpose we can assume that we are working with a finite sample space.

Suppose Ω is a finite sample space. Call the subsets of Ω events. Suppose a probability $P(A)$ is assigned to each event A. Under these assumptions, Kolmogorov's axioms are equivalent to the following slightly long-winded list of axioms:

Axiom 1. For each A, $0 \le P(A) \le 1$.
Axiom 2. If A is impossible, then $P(A) = 0$.
Axiom 3. If A is certain, then $P(A) = 1$.
Axiom 4. If A and B are incompatible, then $P(A \approx B) = P(A)+P(B)$.

Here “A is impossible” means that $A = \emptyset$, “A is certain” means that $A = \Omega$, and “A and B are incompatible” means that $A \cap B = \emptyset$.

We could make this list of axioms more concise. We could omit Axiom 3, for example, because it follows from Axiom 4. But we are interested here in the meaning and justification of the axioms and definitions, not in the most concise way of stating them.

Kolmogorov's axiomatization of probability consists of his axioms together with several definitions. If P(A) > 0, then we call

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

the *conditional probability* of B given A. We say that A and B are *independent* if P(B|A) = P(B). We call a real-valued function X on Ω a *random variable*. We set

$$E(X) = \sum_{i \in \Omega} X(i) \cdot P(\{i\}) \ , \tag{2}$$

and we called E(X) the *expected value* of X.

We will review in turn three standard interpretations of Kolmogorov's axiomatization. The first interpretation takes P(A) to be a person's degree of belief that A will happen. The second takes P(A) to be the frequency with which A happens. The third takes P(A) to be the degree to which the evidence supports A's happening, or the degree to which it is rational to believe that A will happen. Savage (1972) called these the personalistic, objectivistic, and necessary interpretations. They have also been discussed by Nagel (1939), Kyburg and Smokler (1980), Barnett (1982), and many others.

2.1 Belief

The belief interpretation is really a betting interpretation. When a person says her probability for A is 75%, we assume that she will back this up by betting on A and giving 3 to 1 odds. We also assume that she is equally willing to take the other side of such a bet.

Let us review what giving 3 to 1 odds means. It means putting 75¢ on A if the other person puts 25¢ against A. You will lose the 75¢ to the other person if A does not happen, but you will win the other person's 25¢ if A does happen. In effect, you are paying 75¢ for a ticket that returns $1 if A happens. Taking the other side of the bet means paying 25¢ for a ticket that returns $1 if A does not happen.

In a nutshell, then, your probability for A is the price you will pay for a $1 ticket on A. You will pay half as much for a 50¢ ticket on A, and twice as much for a $2 ticket on A.

Why should such prices satisfy Kolmogorov's axioms? And what is the point, if this is what we mean by probability, of defining conditional probability, independence, and expected value in the way Kolmogorov does?

We can argue persuasively for the four axioms. Consider Axiom 4, for example. Suppose A and B are incompatible, your probability for A is 40%, and your probability for B is 20%. Then you would pay 40¢ for a ticket that pays $1 if A happens, and you would pay 20¢ for a ticket that pays $1 if B happens. If you buy both tickets, then in effect you are paying 60¢ for a ticket that pays $1 if $A \approx B$ happens. So your probability for $A \approx B$ must be 60%. Thus $P(A \approx B) = P(A)+P(B)$.

This argument can be elaborated in various ways. One way is to imagine that you post odds for every event and allow another person to choose what bets to make with you at those odds—and which side of the bet to take in each case. In this case, you must satisfy Axioms 1 to 4 in order to keep the person from choosing bets in such a way that she will make money for certain, no matter how the events come out. This is sometimes called the Dutch-book argument.

Conditional probability, in the belief or betting interpretation, is the same as probability—it is probability under new circumstances. Suppose you know that A will happen or fail before B. Then the conditional probability $P(B|A)$ is the degree to which you will believe in B right after A happens—if it happens. In betting terms, it is the amount you will be willing to pay right after A happens for a $1 ticket on B.

This makes it easy to explain the definition of independence. Saying that A and B are independent means that the happening or failing of A will not change your probability for B.

Why should conditional probabilities be given by formula (1)? Suppose your probability for A is 60%, and your probability for B is 50%. Then you are willing to pay 30¢ for a 50¢ ticket on A, and if it does happen, then you are willing to pay the 50¢ you have just won for a $1 ticket on B. If you plan to spend the 50¢ in this way if you win it, then when you pay your 30¢, you are in effect buying a ticket that pays $1 if A and B both happen. Thus your probability for $A \leftrightarrow B$ is 30%. Thus $P(A \cap B) = P(A) \cdot P(B|A)$. This is called the rule of compound probability, and it is essentially equivalent to (1).

We are interpreting the probability $P(A)$ as the amount you are willing to pay for a $1 ticket on A. This ticket pays $1 if A happens and $0 if A does not happen. We can interpret $E(X)$, as given by (2), as the price you are willing to pay for a more complicated ticket X. This ticket pays $X(i), where i is the outcome. If $P(\{i\})$ is your probability for i, then you are willing to pay $[X(i)·P({i})] for a ticket that pays $X(i) if i happens and $0 otherwise. If you buy a ticket like this for each i, then in effect you have bought the ticket X. The amount you have spent is the sum on the right-hand side of (2).

There are some obvious objections to these arguments. First, we may be assuming too much when we assume that a person is willing to set odds on each event and bet on either side of these odds. You can imagine a person being more cautious. She might require you to offer her more than even odds, for example, before she would bet either for or against a particular event. If we allow her to behave in this way, and we still think of the greatest odds she is willing to give on an event as measuring her degree of belief or probability for the event, then we get probabilities that may not add as required by Axiom 4. Such non-additive probabilities have been studied by many authors, including Walley (1991).

The argument for the rule of compound probability also involves some strong assumptions. We assume that the events A and B happen or fail in sequence and that we will know as soon as A happens or fails. We also assume that our probability for B right afterward is well-defined; it does not vary with other circumstances involved in A's happening or failing. Many authors, especially de Finetti (1974, 1975), have tried to relax these assumptions, but then the argument becomes less persuasive (Shafer 1985).

Another point of controversy is whether the belief interpretation is normative or descriptive. Are Kolmogorov's axioms supposed to tell us how a person's degrees of belief *should* fit together? Or are they supposed to describe how people actually behave when given opportunities to bet or when facing other decisions under uncertainty? Most statisticians who subscribe to the belief interpretation say that Kolmogorov's axioms are primarily normative. Whether people conform to these axioms in everyday life is not important to the work of a statistician. Outside statistics, however, the value of the belief interpretation as a descriptive theory is widely debated. Psychologists have given many examples of ways that people do not conform to the axioms in their judgments of probability and in their decisions (Tversky and Kahneman 1986), yet a good deal of modern economic theory assumes that the axioms have some descriptive (or at least predictive) validity (Diamond and Rothschild 1978).

2.2 Frequency

According to the frequency interpretation, the probability of an event is the long-run frequency with which the event occurs in a certain experimental setup or in a certain population. This frequency is a fact about the experimental setup or the population, a fact independent of any person's beliefs.

Suppose we perform a certain experiment n times, under identical conditions, and suppose a certain event A happens k times. Then the relative frequency of A is

$$\frac{k}{n} . \tag{3}$$

Perhaps there is a particular number p toward which this ratio always converges as n increases. If so, then p is the probability of A in the frequency interpretation.

The frequency interpretation is less widely applicable than the belief interpretation. A person can have beliefs about any event, but the frequency interpretation applies only when a well-defined experiment can be repeated and the ratio (3) always converges to the same number.

The frequency interpretation makes Kolmogorov's axioms easy to justify. The axioms obviously hold for the relative frequencies given by (3). The relative frequency of A is always between zero and one. It is zero if A is impossible and never happens, and it is one if A is certain and always happens. If A and B are incompatible events, A happens k_A times, and B happens k_B times, then $A \approx B$ happens k_A+k_B times, and hence the relative frequencies add. The probabilities of the events are the limits of these relative frequencies as n, the number of trials, increases. Because the axioms hold for the relative frequencies, they hold for their limits as well.

The conditional probability of B given A, in the frequency interpretation, is the limit of the relative frequency of B in those trials in which A also happens. Formula (1) follows directly from this definition. Independence means that B happens overall with the same relative frequency as it happens in the trials in which A happens. The expected value of a random variable X is simply the long-run average value of X in many trials.

Many people object to the acknowledged narrow scope of application of the frequency interpretation. Many events for which we would like to have probabilities clearly do not have probabilities in the frequency sense.

We can also question whether the frequency interpretation gives an adequate motivation for the definitions of conditional probability and independence. Why, for example, should we care about the relative frequency of B in the trials where A happens? If we find out that A has happened, then this relative frequency does seem more relevant than the overall relative frequency of B as a guide to whether B will happen. But saying this seems to take us out of the domain of objective frequencies into the domain of belief.

It is also odd, if we begin with frequency as the definition of probability, that we should then expend great effort to prove the law of large numbers—the theorem that the probability of an event will almost certainly be approximated by the event's relative frequency. This was seen as a real problem by the frequentists of the nineteenth century (Porter 1986). But most frequentists nowadays take a more relaxed attitude. Frequency is the definition of probability in practice, they say, but it is convenient in the purely mathematical theory to take probability as a primitive idea and to prove the law of large numbers as a theorem.

Here is a related puzzle. In order to prove the law of large numbers for an event A in our experiment, we must consider a compound experiment, consisting of n trials, say, of the original experiment. We assign probabilities to the possible outcomes of this sequence of experiments, using the probabilities for the original experiment and assuming independence of the trials. Then we choose some number ε, we consider the event B that the relative frequency of the event A in the n trials will be within ε of P(A), and we prove that the probability of B is high. This gives us a frequency interpretation of P(A). But what about the probability of B, and the probabilities of all the other events that can be defined in terms of the n trials? Do they have frequency interpretations? No problem, say many frequentists. We simply consider a yet larger experiment, involving sequences of sequences of trials (Cramér 1946).

The preceding objections have not troubled twentieth-century frequentists, but they have taken a more concrete problem very seriously. This is the problem that probability theory seems to require more than mere convergence of relative frequencies to limits. The convergence must take place at a certain tempo. Yet the frequency interpretation does not impose this. Thus mere frequency does not seem adequate, as a model in the formal sense, for probability theory.

Richard von Mises, in the 1920s and 1930s, proposed that we model probability theory not merely with frequencies but with whole sequences of outcomes. He coined the name "Kollectiv" for a sequence of outcomes whose relative frequencies converge in the manner expected of a random sequence in probability theory. Von Mises's ideas were developed in the 1930s by Jean Ville and Abraham Wald, who showed that it is possible to find sequences of outcomes that satisfy any countable number of the properties that we would expect from a random sequence (Martin-Löf 1969).

During the last three decades, von Mises's ideas have been developed further in terms of the complexity of a sequence, which can be defined as the length of a computer program needed to generate the sequence. A number of mathematicians, including Kolmogorov, have shown that sequences that come close to being maximally complex tend to have the properties we expect from a random sequence (Cover et al. 1989).

2.3 Support

According to the support interpretation, probability is rational degree of belief. The probability P(A) of an event A is the degree to which we should believe A will happen—the degree to which our evidence supports A's happening.

What reason do we have for thinking that there is a precise numerical degree to which our evidence supports A's happening? Twentieth-century proponents of the support interpretation concede that it is difficult to measure degrees of support, but they are convinced that evidence does give support for beliefs. This support may be qualitative rather than quantitative, but it follows certain rules nevertheless, and we can make it quantitative by adopting certain conventions. Kolmogorov's axioms and definitions follow from these qualitative rules and conventions.

One of the most basic qualitative rules advanced by proponents of the support interpretation is the rule that if A and B are incompatible, then the degree of support for A≈B is completely determined by the degrees of support for A and B. To this we may add that it is an increasing function of these two degrees of support; the more support there is for A or for B, the more there is for A≈B. Once we accept these qualitative rules, the numerical rule given by Axiom 4 appears to be a harmless convention (Jeffreys 1961). In fact, it can be derived using a few regularity conditions (Cox 1961, Schrödinger 1947).

Conditional probability and the rule of compound probability can be dealt with similarly. We define conditional probability by saying that P(B|A) is the degree of support for B provided by our present evidence together with the further knowledge that A has happened. We formulate the qualitative rule that the degree of support for A↔B is completely determined by the degree of support for A based on the current evidence along, together with the degree of support for B based on that evidence and knowledge of A. We add that it is an increasing function of both, and we then present the rule of compound probability as a convention or derive it using additional regularity conditions.

It is easy to raise objections to this approach. To begin with, we can question the qualitative rules. Why should the degree of support P(A≈B) depend only on the degrees of support P(A) and P(B), and not on other aspects of these two events or other aspects of the evidence? There does not seem to be any argument for this qualitative principle, aside from the fact that the familiar numerical rule satisfies it. In some alternative theories (e.g., Shafer 1976), the principle is not satisfied.

Even if we accept the existence of well-defined degrees of support based on our current evidence, we can question whether conditional degrees of support exist. Because we might learn that A is true in many different ways, it may not be appropriate to talk without qualification about the support for B based on our current evidence together with knowledge of A (Shafer 1985).

3 The Three Interpretations in Practice

We have been discussing the three interpretations of probability as if they were completely unrelated—as if Kolmogorov's axioms and definitions were all they had in common. In fact, the three interpretations are thoroughly entangled. They are entangled historically, conceptually, and practically. In this section, we look at the entanglement in statistical practice.

In statistical practice, proponents of all three interpretations are interested in both frequencies and degrees of belief. All three groups make inferences about frequencies, and they all use probabilities, in one way or another, to express these inferences. There is disagreement about how to make inferences; proponents of the belief interpretation and the support interpretation use Bayesian methods, while proponents of the frequency interpretation use sampling-theory methods. But in both cases, the inferences are about frequentist probabilities. There is also disagreement about how to express the inferences; sampling-theory methods express them with probabilities with subtle frequency interpretations, while Bayesian methods express them with probabilities that are labelled outright as degrees of belief or degrees of support. But in practice even the probabilities produced by sampling-theory methods are interpreted as degrees of belief.

Bayesian and sampling-theory inference are discussed in more detail in Chapters 1 and 7 of this volume. Discussions that emphasize the differences between the two approaches include Efron (1978) and Barnett (1982).

3.1 Bayesian Inference

Suppose we flip a coin 10 times, and we get the sequence HHTTHTHHHH; 7 heads and 3 tails altogether. What should we think about the true probability of heads?

If we write p for the true probability of heads, then the probability of the sequence HHTTHTHHHH is

$$p^7(1-p)^3. \qquad (4)$$

This is graphed as a function of p in Figure 1. It seems reasonable to take this function as a measure of how much we should believe different values of p. Thomas Bayes and Pierre-Simon Laplace, two eighteenth-century students of probability, suggested that we use it to find probabilities for p. We multiply (4) by a constant so that it will integrate to one, and we use the result as a probability density:

$$f(p) = 1320\, p^7(1-p)^3. \tag{5}$$

Using this probability density, we can give probabilities for any interval of values for p. The probability of p being between 0.54 and 0.86, for example, is 77%.

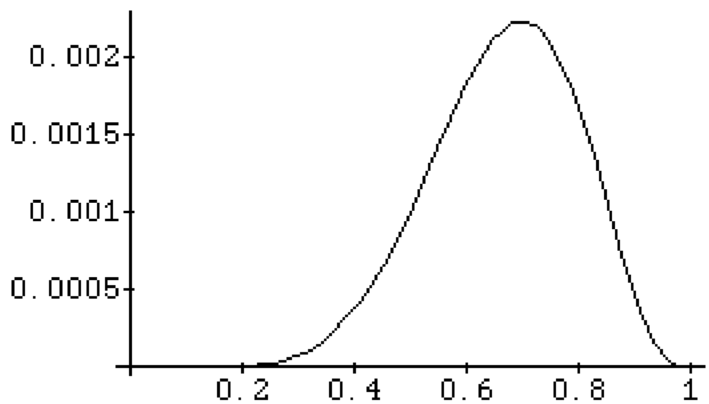

Figure 1. The likelihood function for p resulting from 7 heads and 3 tails

The probability density given by (5) is called the “posterior” density for p. The function given by (4) is called the “likelihood function.” We have simply multiplied (4) by a constant to get (5), but we can incorporate into the process a probability density g(p) based on other evidence. We call g(p) the “prior” density, and we take the posterior density to be proportional to the product of g(p) and the likelihood function. Thus formula (5) gives the posterior density for our problem only in the case where the prior g(p) is uniform (i.e., where g(p) =1 for all p, $0 \le p < 1$).

This approach to statistical inference was made popular by Laplace, and in the nineteenth century it was called the method of inverse probability. Today we call it Bayesian inference, and we base it on what we anachronistically call Bayes's theorem. Bayes's theorem says that if $A_1, A_2, \ldots, A_n$ are incompatible hypotheses, one of which must be true, then

$$P(A_i|B) = K\, P(A_i) \cdot P(B|A_i),$$

where K is a constant in that it does not depend on A. Here $P(A_i|B)$ is the posterior, $P(A_i)$ is the prior, and $P(B|A_i)$ is the likelihood. This theorem is easy to prove if we accept Kolmogorov's axioms as a starting point, but it is conceptually troublesome, because it involves conditional probabilities in two directions (the probability of A_i given B and the probability of B given A_i), whereas the justification of conditional probability that we reviewed in Section 2.1 is based on a single sequence of events, with the assumption that both the events themselves and our knowledge of them unfold together in that sequence.

Both the support and the belief interpretations use Bayesian inference. The difference between them is in their interpretation of the prior probabilities. Proponents of the belief interpretation regard prior probabilities as personal beliefs. Proponents of the support interpretation try to find objective grounds for choosing a prior distribution. In the case of the coin, for example, they regard the uniform prior density as an expression of ignorance. The support interpretation is rejected by most statisticians, because the case for objective priors is a confused one. But the belief interpretation has its own problem here. This lies in the apparently objective and frequentist nature of the true probability p. Are we giving probability a belief interpretation if we interpret the prior and posterior probabilities as beliefs but interpret the “true unknown probability p” itself as an objective property of the coin?

De Finetti (1974, 1975) has argued that the apparently frequentist p is merely a way of talking; behind it lie purely subjective ideas about the symmetry of our beliefs about a long sequence of successive flips of the coin. Nonetheless, the workaday world of Bayesian statistics seems to accept a dual interpretation of probability. In practice, Bayesians accept models that hypothesize frequentist probabilities. They differ from the frequentists only in that they use probabilities interpreted as beliefs in order to make inferences about the probabilities interpreted as frequencies.

3.2 Sampling-Theory Inference

Consider again the problem of making judgments about the true probability p after observing 10 tosses. The frequentist approach considers the frequency properties of different ways of making such judgments.

Suppose we write X for the number of heads in 10 tosses, and we say that we are going to estimate p by $\frac{X}{10}$. The expected value of this estimator is p, and its standard deviation is $\sqrt{\frac{p(1-p)}{10}}$, which is equal at most to 0.16. By the central limit theorem, we expect that the estimator will be within one standard deviation of p about 68% of the time. So if we say that p will be in an interval that extends 0.16 on either side of $\frac{X}{10}$, we will be right at least 68% of the time. When X falls equal to 7, this interval is from 0.54 to 0.86. So we call the interval from 0.54 to 0.86 a 68% confidence interval for p.

Textbook expositions of this method keep "confidence" quite distinct from "probability." They emphasize, moreover, that the confidence coefficient of 68% is ultimately a frequentist probability: it is approximately the frequency with which a certain method produces an interval that covers p. Yet the language is designed to encourage us to interpret this frequentist probability as an opinion about p. It is the degree to which we can be confident that p is between 0.54 and 0.86. Most users of statistics see little difference between this and a Bayesian degree of belief.

Confidence intervals are only one method in the repertoire of the frequentist statistician. Another important method is statistical testing, especially the use of goodness-of-fit tests for statistical models. We need not describe such tests here; they are discussed in most statistics textbooks. But they too produce frequentist probabilities (significance levels or P-values) that are given a belief interpretation at the level of practice (Box 1980).

4 Three Aspects of Probability

Probability is a complex idea. Belief, frequency, and support are three of its aspects, and it has other aspects as well. One way to bring together the many aspects of probability is to emphasize the special situation that occurs when we repeatedly perform an experiment for which we know only the long-run frequencies of outcomes. In this special situation, we know the frequencies, and we know nothing else that can help us predict the outcomes. The frequencies therefore determine odds, or prices for tickets on events. These are more than personal prices; they are fair prices, in the sense that they break even in the long run. Because the frequencies are our only evidence, they also determine well-defined numerical degrees of support for events, or degrees to which it is warranted or rational to believe that the events will happen.

The triangle in Figure 2 symbolizes how the ideas of fair price, warranted belief, and knowledge of the long run hold together, both conceptually and historically. Conceptually, we can start at any point in the triangle and reason in the direction of the arrows. Historically, probability began as a theory of fair odds in games of chance, and ideas of probability (which then meant warranted belief) and frequency were only gradually incorporated into the theory.

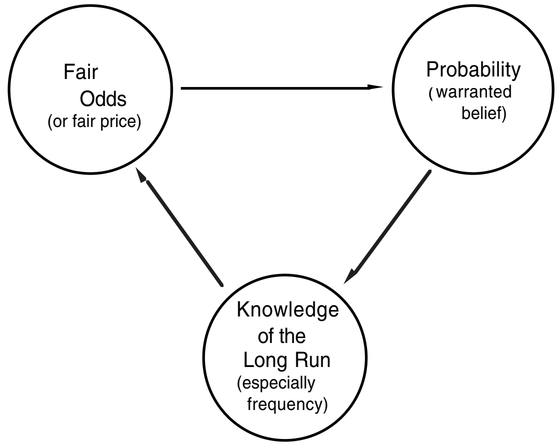

Figure 2. The triangle of probability ideas

In the first part of this section, we use the triangle of Figure 2 to gain a clearer understanding of why the different aspects of probability are aspects of a single concept. Then we use the triangle to review the history of the conceptual development of probability. We conclude with some suggestions for recasting the standard interpretations so as to regain the unity represented by the triangle.

Shafer (1991a) develops these themes further. Shafer (1990) describes the conceptual triangle in more detail. Daston (1988) and Shafer (1991b) discuss aspects of the historical development.

4.1 The Conceptual Triangle

Let us consider how we can move around the triangle conceptually, starting with our knowledge of the long run.

The knowledge of the long-run that we have in the special situation described by probability theory is quite extensive. We know long-run frequencies for the outcomes of our experiment. We also know about the rate at which these frequencies are likely to converge, and we know betting schemes are futile. We know we cannot accomplish anything by strategies for compounding bets on successive events at the odds given by the long-run frequencies. No such strategy can assure us of a net gain or give us any reasonable expectation of substantially multiplying our initial stake.

This knowledge of the long-run already refers to odds for events in individual experiments. These odds are fair because we break even in the long run by betting at them. By compounding bets, we can derive fair odds for events that involve more than one experiment, and we can study how these odds change as the experiments are performed. This is the arrow upward and to the left in Figure 2, the arrow from knowledge of the long run to fair odds on all events.

Once we have fair odds, or fair prices for tickets on events, we can use these odds or prices as degrees of belief. Because the odds are fair odds, not just personal odds, the degrees of belief are warranted degrees of belief, not just personal degrees of belief. From the properties of the fair odds, we can derive rules for these warranted degrees of belief, which we may call probabilities. This is the arrow to the right in Figure 2. The rules we derive for probabilities are similar to Kolmogorov's axioms and definitions, except that they involve probabilities changing as the experiments are performed, not probabilities conditional on arbitrary events.

From the rules for probabilities, we can deduce the knowledge of the long run with which we began. This is the arrow downward and to the left in Figure 2.

Any of the three circles in Figure 2 can be taken as a starting point for the mathematical theory of probability. The theory of algorithmic complexity theory starts with knowledge of the long run. Kolmogorov's axioms start with warranted belief. Similar axioms have been formulated for fair price.

The fact that knowledge of the long run, fair price, and warranted belief can each be used as a starting point for the mathematical theory does not mean that any one of these ideas is sufficient for grounding probability in a conceptual sense. The axioms or assumptions we need in order to begin with any one of these starting points can be understood and justified only by reference to the other aspects of the picture. The three aspects of probability are inextricably intertwined.

4.2 The Historical Triangle

The historical development of mathematical probability followed Figure 2, but with fair price as the starting point. The theory began with problems of equity in games of chance, and it only gradually expanded to encompass the ideas of probability and knowledge of the long run.

What we now call mathematical probability began in the 1650s with the work of Blaise Pascal and Pierre Fermat. They were primarily interested in equity—in finding fair odds in games of chance. They did not discuss probability, or the weighing of arguments, which was an important topic at the time. Both probability and frequency were brought into the theory later, by James Bernoulli. In his *Ars conjectandi*, published in 1713, Bernoulli explained that probability is degree of certainty, and he related certainty to equity by saying that an argument has a certain share of certainty as its fair price. He also brought frequency into the theory, by proving the law of large numbers.

Bernoulli's moves from equity to degree of certainty and then to frequency are represented by two of the arrows in Figure 2. The third arrow, from frequency back to equity, came much later. Today we are accustomed to saying that the odds given by probability theory are fair because they are odds at which we will break even in the long run. More generally, the expected value E(X) of a random variable X is the fair price of X because it is the price at which we will break even in the long run. This idea appears very late in the probability literature, however. It was first formulated, apparently, by Condorcet in the 1780s (Todhunter 1865), and it did not become popular until the nineteenth century.

The weight of opinion on the foundations of probability theory moved around the triangle even more slowly. Ideas of fairness remained at the foundation of the theory well past 1750. It was only as the probabilistic theory of errors became important in the second half of the eighteenth century that probability, in the sense of rational belief, became fully independent of ideas of equity. An important signpost in this development was Laplace's influential *Théorie analytique des probabilités*, first published in 1812. Laplace interpreted probability as rational degree of belief, and he took the rules for probability to be self-evident. He did not derive them, as his predecessors had, from rules of equity.

In retrospect, Laplace's views look much like the support interpretation, but he did not make the kind of distinction between support, belief, and frequency that we make today. Though he began with the idea of support or rational degree of belief, he did not hesitate to follow Bernoulli in deducing that the long-run frequencies of outcomes will approximate their probabilities.

The frequency interpretation arose in the nineteenth century because of the influence of empiricist philosophy. The empiricists saw both fairness, degree of certainty, and rational belief as metaphysical ideas, ideas not grounded in reality. They saw frequency as the only empirical grounding for the theory. So probability should start with frequency. The mathematicians should not pretend, as Bernoulli and Laplace had, to derive facts about frequency from metaphysical ideas about subjective certainty or rational belief.

The frequency interpretation became dominant only in the late nineteenth and early twentieth centuries. As its own shortcomings became evident, twentieth-century scholars sought new foundations for the older non-frequentist idea of probability. This produced the belief and support interpretations. The belief interpretation, first advanced by F.P. Ramsey and Bruno de Finetti in the 1920s, went back to the ideas of the pioneers, except that it replaced fair odds and rational degree of belief with personal odds and personal degree of belief. The odds were odds at which a particular person would bet, not odds at which it was fair to bet. This made the interpretation empirically respectable. A person's betting behavior is an empirical fact, not a metaphysical idea like fairness. The support interpretation was less of a departure and more of a continuing defense of Laplace's ideas against the empiricism of the frequentists. John Maynard Keynes and Harold Jeffreys were very influential in this defense.

4.3 Unifying the Standard Interpretations

The historical and practical entanglement of the standard interpretations, together with their unity in the special situation of a sequence of experiments for which we know long-run frequencies, suggests that they should be recast in a way that emphasizes their commonalities.

For the belief interpretation, this would involve returning to probability's original emphasis on *fair* odds rather than *personal* odds. Ramsey and de Finetti's attempt to drop fairness was a mistake. There is no reason for a person to have personal odds at which she would bet on either side. But a person can draw an analogy between her evidence and the special situation where fair odds are known. She can say that her evidence is analogous, in its strength and import, to knowing certain fair odds, which are based on long-run frequencies. This recasting of the belief interpretation pulls it toward both the frequency and support interpretations.

We also need to acknowledge the subjective aspects of the frequency story. A full account must go beyond the existence of frequencies to the fact that we know these frequencies and nothing more that can help us predict. The randomness of a sequence is not an objective fact about the sequence in itself. It is a fact about the relation between the sequence and the knowledge of a person. This point emerges in various ways in the frequentist foundations pioneered by von Mises and Kolmogorov. In Kolmogorov's complexity theory, for example, the complexity of a sequence is defined in terms of the length of the computer program needed to generate it, and this depends on what programming language is used. This means that what is random for a person using one programming language may not be so random for another person. Frequentists tend to minimize this non-objective aspect of the complexity idea by talking about longer and longer sequences—or even by taking refuge in the idealization of infinite sequences (Uspenskii et al. 1990). But if we refuse to minimize it, we create another point of contact between frequentist and belief foundations.

We can build on this point of contact by emphasizing the ordering of events when we explain the belief interpretation of conditional probability. If we begin with an ordering of events, we have a sequence of events, and hence frequencies have a place within the belief interpretation.

With this approach, the three interpretations begin to resemble each other. All three are really about the special situation where we have a sequence of experiments with known chances. Most applications of probability, including statistical inference, lie outside this special situation, of course. But we can think of these applications as various ways of relating real problems to the special situation. Much standard statistical modelling amounts to using the special situation as a standard of comparison. Statistical arguments based on sampling or randomization depend on artificially generated random numbers which simulate the special situation. Bayesian analyses are arguments by analogy to the special situation.

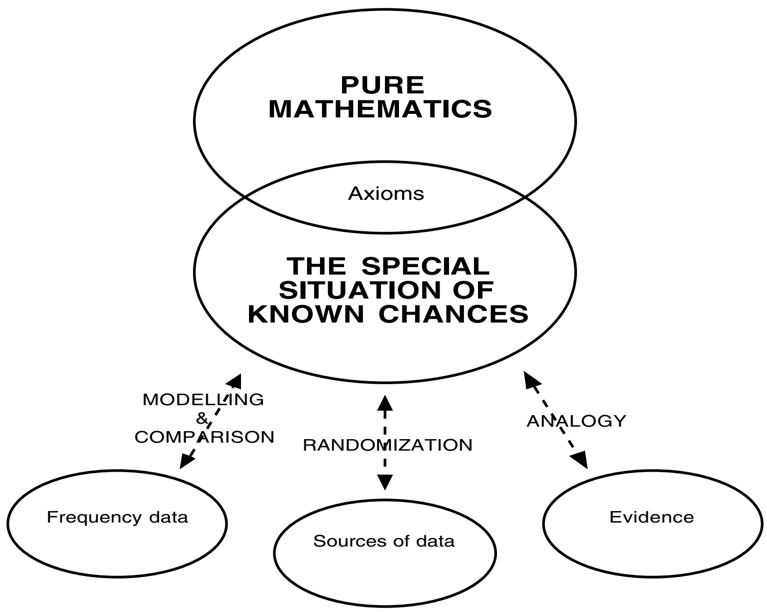

Figure 3. A unified interpretation of probability

Figure 3 summarizes this approach to the meaning of probability. Here probability becomes a story about a special situation involving known long-run frequencies. Various aspects of this story can be made into pure mathematics, and we can base this pure mathematics on Kolmogorov's axioms. The different applications of probability do not, however, depend on different interpretations of the axioms. Instead, they are different ways of using the probability story.

5 Lessons for Teaching

The main message of this chapter is that probability is a complex idea. It is not simply a set of axioms, nor is it a single interpretation of these axioms. It is a tangle of ideas that took hundreds of years to evolve.

This complexity is evident in textbooks on probability and statistics. A few textbooks manage to take an uncompromising ideological line, either frequentist or Bayesian, but this is hard to sustain. We need to appeal to all the aspects of probability in order to teach the mathematics of probability effectively. We must appeal to frequency in order to explain why probabilities add and in order to explain the significance of the expected value of a random variable. We must appeal to belief when explaining the idea of conditional probability. We must appeal to support when explaining why scientists want to use probabilistic ideas in data analysis.

An understanding of the complexity of probability should teach us humility when teaching the subject. We should be wary of pointing to any particular aspect of probability and saying, "This is what it really means." In particular, we should be wary of telling students that probability is simply a branch of pure mathematics. Probability is not measure theory. It did not begin with Kolmogorov.

The complexity of probability should also make us wary of any strict ideology in teaching statistics. Most elementary textbooks take a sampling-theory viewpoint, but they do not adhere to it strictly, and there are good reasons for this laxness. Some textbooks take a Bayesian approach; here too the teacher needs to be aware that there are good reasons why the proclaimed subjective interpretation is carried only so far.

The ways in which probability are used, in statistical inference and elsewhere, are varied, and they are always open to criticism. We should guard, however, against the idea that a correct understanding of probability can tell us which of these applications are correct and which are misguided. It is easy to become a strict frequentist—or a strict Bayesian—and to denounce the stumbling practical efforts of statisticians of a different persuasion. But our students deserve a fair look at all the applications of probability.

Acknowledgments

Research for this article has been partially supported by the National Science Foundation through grant IRI8902444 to the University of Kansas.

References

Barnett, V. (1982), *Comparative Statistical Inference*, Second Edition, New York: John Wiley.

Box, G .E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society,* Ser. A, 143, 383-430.

Cover, T. M., Gacs, P., and Gray, R. M. (1989), "Kolmogorov's Contributions to Information Theory and Algorithmic Complexity," *The Annals of Probability* **17** 840-865.

Cox, R. T. (1961), *The Algebra of Probable Inference*, Baltimore: The Johns Hopkins Press.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton New Jersey: Princeton University Press.

Daston, L. (1988), *Classical Probability in the Enlightenment,* Princeton New Jersey: Princeton University Press.

Diamond, P., and Rothschild, M. (1978), *Uncertainty in Economics*, New York: Academic Press.

Efron, B. (1978), "Controversies in the Foundations of Statistics," *American Mathematical Monthly*, 85, 231-246.

Finetti, B. de (1974, 1975), *Theory of Probability*, 2 vols., New York: Wiley.

Jeffreys, H. (1961), *Theory of Probability*, Third Edition, Oxford: Oxford University Press.

Kyburg, H. E., Jr., and Smokler, H. E., eds. (1980), *Studies in Subjective Probability*, Second Edition, New York: Robert E. Krieger.

Martin-Löf, P. (1969), "The Literature on von Mises' Kollektivs Revisited," *Theoria*, 35:1, 12-37.

Nagel, E. (1939), *Principles of the Theory of Probability* (Volume 1, Number 6 of the *International Encyclopedia of Unified Science*), Chicago: University of Chicago Press.

Porter, T. M. (1986), *The Rise of Statistical Thinking, 1820-1900*, Princeton New Jersey: Princeton University Press.

Savage, L. J. (1972), *The Foundations of Statistics*, Second Edition, New York: Dover.

Schrödinger, E. (1947), "The Foundation of Probability—I," *Proceedings of the Royal Irish Academy,* Ser. A, 51, 51-66.

Shafer, G. (1976), *A Mathematical Theory of Evidence,* Princeton New Jersey: Princeton University Press.

Shafer, G. (1985), "Conditional Probability" (with discussion), *International Statistical Review*, 53, 261-277.

Shafer, G. (1990), "The Unity of Probability," in *Acting Under Uncertainty: Multidisciplinary Conceptions*, ed. G. von Furstenberg, New York: Kluwer, pp. 95-126.

Shafer, G. (1991a), "Can the Various Meanings of Probability be Reconciled?" To appear in *Methodological and Quantitative Issues in the Analysis of Psychological Data*, Second Edition, ed. G. Keren and C. Lewis, Hillsdale, New Jersey: Lawrence Erlbaum.

Shafer, G. (1991b), "The Early Development of Mathematical Probability." To appear in *Encyclopedia of the History and Philosophy of the Mathematical Sciences*, ed. I. Grattan-Guinness, London: Routledge.

Todhunter, I. (1865), *A History of the Mathematical Theory of Probability*, London: Macmillan.

Tversky, A., and Kahneman, D. (1986), "Rational Choice and the Framing of Decisions," *Journal of Business*, 59, S251-S278.

Uspenskii, V.A., Semenov, A.L., and Shen', A.Kh. (1990), "Can an Individual Sequence of Zeros and Ones be Random?" *Russian Mathematical Surveys*, 45, 121-189.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall.