

The Subjective Aspect of Probability

Glenn Shafer

Graduate School of Management

Rutgers University

72 New Street

Newark, New Jersey 07102-1895

shafer@draco.rutgers.edu

Abstract:

Subjectivity is an integral aspect of all applications of probability. This chapter demonstrates this by showing how the unified informal story of probability, in which a spectator's beliefs in certain events match their frequencies, can be used to understand elementary examples of statistical testing.

Is subjective probability a kind of probability, corresponding to a particular interpretation of the mathematical calculus of probability? Or is subjectivity always an integral aspect of probability, even in applications such as statistical testing, where the objective aspects of probability are usually emphasized? In this chapter, I argue that subjectivity is an aspect of all applications of probability. When we enunciate clearly the subjective aspects of supposedly objectivistic applications, the subjectivist critique of these applications loses its force. It is not necessary that these applications be rejected or be replaced with more complicated Bayesian procedures. It is only necessary that they be properly understood.

When we learn the mathematics of probability, we learn an informal story in which belief and frequency are unified. This story has many variations, but it usually involves a sequence of experiments in which known odds simultaneously define fair prices, warranted degrees of belief, and long-run frequencies. Different ways of using probability are understood most clearly when seen as different ways of using this informal story. Thus subjectivity enters into probability in two ways. First, subjectivity is part of the informal story itself. The probabilities in the story are, *inter alia*, the beliefs of some person, real or imaginary. Second, it is up to us to bring the informal story to bear on a practical problem. In doing so, we construct an argument, which must itself be criticized and subjectively evaluated.

In previous essays, I have described the unified informal story of probability and argued for its primacy over any particular axiomatization of probability. I have also made the general point that different applications of probability use the informal story in different ways. Here I review and refine these arguments with emphasis on a particular class of applications: statistical tests. In many cases, as

we shall see, statistical tests use instances of the informal story simply as standards against which to rate the performance of a forecaster or method of prediction. This is very different from using the informal story as a representation (model or map) of a problem. By saying this clearly, we can dispel much of the confusion and controversy that now besets statistical testing.

The larger point of this chapter is that proponents of subjective probability can afford to recognize the diversity of ways in which the informal story of probability can be used. Most frequentists, deeply influenced by the empiricism of the late nineteenth and early twentieth centuries, consider an application of mathematical probability legitimate only if each probability number is mapped to an empirical frequency. Despite the anti-realism of de Finetti, subjectivists have tended to adopt an equally rigid understanding of the relation between theory and application: an application is legitimate only if each probability number is mapped to a belief or betting rate (actual or perhaps only proposed) about a practical question. This foundational rigidity may have been helpful when subjectivists had few practical Bayesian applications to their credit, but it is not necessary today. The self confidence of today's subjectivists should allow them to lay claim to the subjective nature and legitimacy of all uses of probability.

This chapter is divided into two sections. The first section reviews the argument for the unified understanding of the informal story of probability. The second section relates this story to some simple examples of statistical testing.

1. The Unified Informal Story

Subjectivists and frequentists each have their own informal stories about probability, stories that they take to underly and justify the formal theory. The subjectivist story is about the betting rates of ideal rational agents, while the

frequentist story is about the properties of exceptionally complex and unpredictable (i.e., random) sequences. The informal story I have in mind combines the subjectivist and frequentist stories. It involves both a sequence and a person who has a certain limited kind of knowledge about the sequence. This unified story is familiar in its basics; we learn it inadvertently when our teachers slide from back and forth between subjectivist and frequentist ideas in order to persuade us to accept the various rules of probability. But it has not received much philosophical attention. Those who could give it such attention have usually chosen instead to defend one of the narrower stories.

In order to understand the unified informal story fully, we must first describe it in its own terms and then relate it to its various axiomatizations, each of which captures or emphasizes only certain of its aspects. I have made a beginning on these tasks in earlier essays.¹ There is not enough space here to discuss axiomatizations, but I will briefly recount the story and explain why I prefer it to the narrower stories.

1.1. A Brief Recounting of the Story

Since it must capture the frequency aspects of probability, an adequate recounting of the unified informal story must have some representation for a sequence of events. The simplest and perhaps oldest such representation is the event tree.² Figure 1 is an example. As we see in this figure, the events in an event tree result from a sequence of experiments, and the experiment performed in a given situation may depend on what has happened so far. The figures use circles for situations in which an experiment is performed and octagons (stop signs) for situations in which experimentation has stopped.

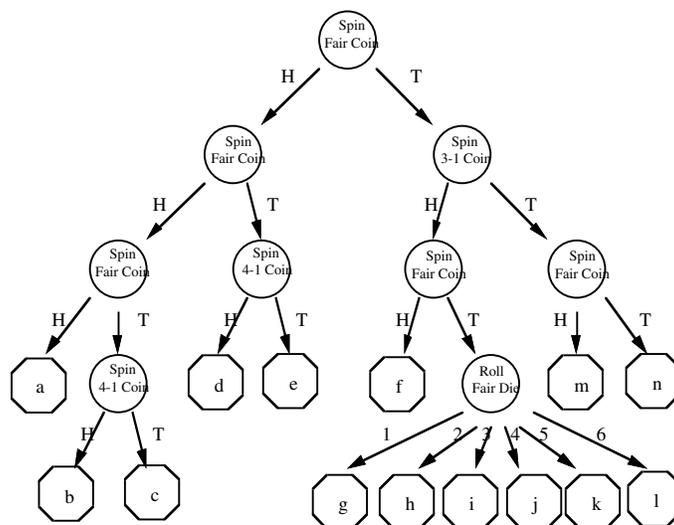


Figure 1. An event tree.

The unified informal story also involves a spectator, who observes the outcome of each experiment as it is performed. This spectator begins with some limited knowledge about how the experiments will turn out; she can make certain predictions about what will happen on average, but she cannot go beyond this to predict reliably the outcomes of individual experiments.

Each experiment has several possible outcomes, and a probability is specified for each outcome. These probabilities have several roles. They define fair odds, warranted degrees of belief, and long-run frequencies. The odds corresponding to the probabilities are fair because the spectator knows that if she makes many small bets at these odds—say a small bet on the outcome of each experiment as she moves down the tree—she will approximately break even. She also knows that she has no way of finding a strategy for betting at these odds that can give her any reasonable expectation of substantially multiplying her initial stake. Since she is willing to bet at these odds, the probabilities may be considered her degrees of belief, and since the odds are fair, her degrees of belief may be considered warranted. Finally, in a limited way, she interprets the probabilities as

frequencies: she knows that if she bets on the outcome of each successive experiment, the frequency with which she wins will approximately equal, in the long run, the average of the probabilities for the outcomes on which she bets. (Notice that this “frequency interpretation” does not involve repeatedly going down the tree. It refers to the spectator's single trip down the tree. It is only an interpretation of certain average probabilities, however; it is not an interpretation of each and every probability in the tree.)

In order for our assertion about the spectator breaking even to be reasonably accurate, every path down the tree must go through many (a few hundred at least) situations before coming to a stop sign, and the spectator must specify a complete strategy for laying bets. For each situation, she must specify how she will, if she arrives in that situation, bet on the experiment performed there, subject to the constraint that she will have the money to pay off the bet. (How much she has in the situation is determined by her initial stake together with her strategy, for the strategy determines what she will win and lose on the way down to the situation.) When we say the spectator will approximately break even, we mean that she will approximately break even no matter what path she takes down the tree and what strategy she chooses. After she has gone down the tree, she will see ways she could have laid her bets so as to win heavily, but she has no way of choosing such a strategy in advance, and she is practically certain that any strategy she does choose will be of no avail.

In addition to the outcomes of individual experiments, the spectator can also bet on events involving more than one trial. In Figure 1, for example, she can bet on the event that the path down the tree will end up in the set {a,d,e}, and this event may depend on three different experiments, those performed in the situations labelled U, V, and W in Figure 2. (If the spin³ in U yields tails, the event fails. If

it yields heads, then we move on to the spin in V. If the spin in V yields tails, the event happens; if it yields heads, we move on to the spin in W. If the spin in W yields heads, the event happens; if it yields tails, the event fails.) In general, any set of stop signs is an event, and a bet on any such event can be compounded from bets on individual experiments, so that its fair price is determined by the fair prices for bets on the individual experiments. In other words, the probabilities for the individual experiments determine probabilities for all events in the tree—probabilities for all sets of stop signs.

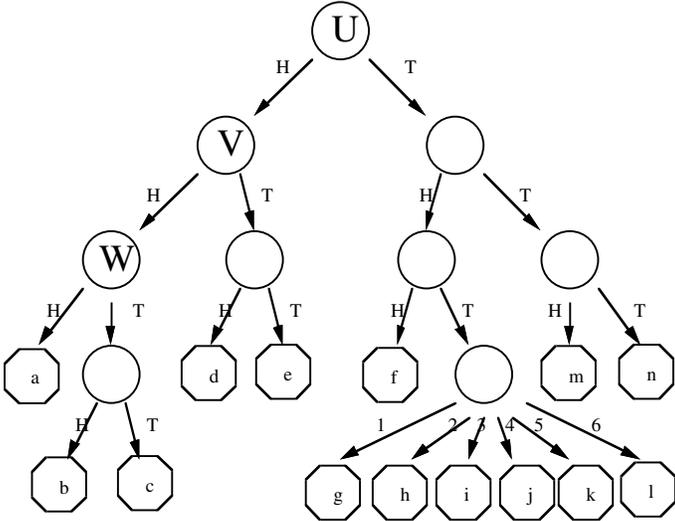


Figure 2. The event {a,d,e} depends on experiments in U, V, and W.

The spectator’s probabilities change as events move down the tree. Her knowledge unfolds with events; she sees the outcome of each experiment as it is performed. So as she moves on to the next situation, she changes her probabilities for the experiment she just saw performed, giving probability one to the outcome she actually observed. Since more complicated events are compounded from events involving the individual experiments, she also changes her probabilities for them as she moves down the tree. So when we speak about the spectator’s

probabilities, we must, in general, specify the situation to which we are referring—the situation in which she has those probabilities. When we talk about the probabilities for the outcomes of an experiment performed in a given situation, we usually mean the probabilities in that situation. But in general, we can talk about the probability for any event in any situation.

Among the events in the tree are events that correspond to the assertion that a given strategy will approximately break even. Thus this assertion itself has a probability. In the initial situation, before any experiments are performed, this probability is close to one, expressing what we have described as the spectator's knowledge or practical certainty that she will approximately break even. We can similarly express her practical certainty that she cannot substantially multiply her initial stake: her probability in the initial situation that a given strategy will multiply her stake by k or more is never more than $1/k$.

It is part of the story that these practical certainties match realities in the spectator's situation. The story is about more than the spectator's inner life. According to the story, she really does move down a tree of experiments, and her ability to predict the outcomes really is limited. She really is unable to pick out a winning strategy. Any strategy that she does choose for placing small bets on successive experiments really will approximately break even. The story is a story about knowledge—a story about the relation between fact and belief.

1.2. Why This Story?

Why should we be interested in this unified story? Why not instead base our understanding of probability and its applications on the separate but narrower stories of the subjectivists and the objectivists?

The shortcomings of the objectivistic story have been exhaustively discussed during the past several decades. Here let me simply point out that these shortcomings lie not in the coherence of the story itself, but in the difficulty of applying it to a broad range of practical problems. Indeed, proponents of the objectivistic story are usually outspoken about the need to restrict application. Some argue that probability should only be used in cases where data is generated by random mechanisms (Freedman et al. 1991). Others find objectivity in the mathematical theory of infinite sequences and leave us to puzzle over how application to finite problems can ever be justified.

Criticisms of the subjectivistic story also center on the difficulty of using it. It is argued that we often have inadequate information on which to base the betting rates that would make us like the ideal rational agents in the story. My own interest in the theory of belief functions, which uses non-additive numerical degrees of belief (Shafer 1990b) has encouraged me to push the criticism one step further: it is only in the unified story that we have grounds for calling our betting rates fair and hence using them both for buying and selling.

The standard expositions of the subjectivistic story do not place event trees in the foundation of the theory. Sequences of events are seen merely as one thing about which we can have beliefs. But it turns out that sequences of events are needed in order to justify the idea of belief change by conditional probability; without the “protocol” for new information represented by an event tree, we are led into paradox (Shafer 1985). Thus even the internal logic of the subjectivistic story pushes it in the direction of the unified story for which I am arguing (Dawid 1982).

My purpose in this chapter is to show that there are practical reasons for favoring the unified story that go beyond these general arguments. There are some

applications of probability that can be understood in terms of the unified story but not in terms of the narrower stories.

2. Evaluating Categorical Predictions

Statistical testing, presented with little theory, is often very persuasive. If a forecaster does no better than chance, there is no reason to hire her. When one treatment does better than another no more often than might be expected by chance, its performance provides no evidence that it is better. When an additional variable improves the performance of a prediction equation no more than might be expected by chance, the improvement is a poor argument for adding it to the equation.

When we turn to theoretical accounts of testing, on the other hand, we find confusion and controversy. Every teacher of elementary statistics knows how confused students are by the objectivistic accounts we teach, and every theoretical statistician is familiar with the mockery these accounts evoke from subjectivists and other skeptics.⁴ Bayesian elaborations of the objectivistic accounts are also controversial; they add to the complexity of the objectivistic accounts and correct only some of their shortcomings.

Why is it so difficult to make theoretical sense of statistical testing? The difficulty, I believe, lies in an unspoken but powerful assumption about how probability theory should be related to practice. We assume, without reflection, that any probability model we formulate to study a phenomenon must be a model for—a representation of—that phenomenon. So when we undertake to explain a statistical test (or rather, to improve the apparently shallow explanation we first found persuasive), we begin by trying to construe the probability model involved in the test as a representation of the phenomenon being tested. We try to make the

informal story corresponding to the model a story about that phenomenon—a story about the behavior of the forecaster or what she is forecasting, a story about the effect of the treatment, or a story about the effect of the additional variable. We forget that the model and the story originally stood apart from the forecaster, treatment, or variable, as an independent standard to which to compare their performance.

The unified informal story of probability can help us keep our hands on the knowledge that testing involves comparison rather than representation. This unified story can serve as a clear standard for comparison in a way that its objectivistic and subjectivistic cousins cannot, for within the unified story there is a spectator with clearly delimited powers of prediction, and it is to this spectator that we compare our forecaster or our prediction equation.

Any particular statistical test involves, of course, a particular story; we are compare our forecaster not with the unified informal story in general but with a particular instance of it, an instance with a particular event tree and particular probabilities. For brevity, let us call an instance of the unified informal story a “stochastic story.” For clarity, let us reserve the name “forecaster” for the real forecaster we wish to evaluate (as opposed to the “spectator” in the stochastic story), whether it be a person, a prediction equation, or an expert system. (We may speak of forecasting or prediction even when we are dealing with assertions about the past or present.⁵ We require only that after the forecaster makes her prediction we are able to classify it as right or wrong.)

We deliberately construct the stochastic story that serves as our standard for comparison. We often construct several. We may begin by comparing the forecaster's performance to what can be achieved by a nearly clueless spectator in a very austere stochastic story. If the forecaster can do better than this spectator,

then we may move on to a stochastic story whose spectator is more (or perhaps merely differently) advantaged. Continuing in this way if necessary, we may (or may not) find a stochastic story in which the performance of the spectator roughly matches the performance of our forecaster. But none of this requires us to go beyond the idea of rating the forecaster's performance. At no point are we required to think of the forecaster herself or of the phenomenon she is forecasting as part of a stochastic story.

There are some general principles that can guide our search for an appropriate stochastic story. We must make the story and the situation of the forecaster comparable without contriving to force any particular conclusion. No general principle can guarantee, however, that the comparison with the stochastic story will be persuasive. In the end, this comparison is only an argument, and like any other nondemonstrative argument, it is open to criticism and counterargument. A particular stochastic story will not be persuasive unless equally natural stochastic stories give similar or consistent results.

I will discuss two simple examples of categorical prediction. In both examples, as we will see, the success of the prediction can be evaluated by comparison with a stochastic story. The two evaluations can be extended to deal with problems that are usually treated as statistical testing problems. The first corresponds to testing whether a binomial parameter is equal to $\frac{1}{2}$, and the second corresponds to testing independence in a 2x2 table.

The analysis of these simple examples falls short of a general theory of statistical testing. But there are some obvious ways to extend the analysis. In order to extend it to the kinds of problems that are usually treated by goodness of fit tests and tests of independence in larger tables, we will need to use the ideas on the evaluation of probability forecasting developed by Dawid (1984, 1985,

1986, 1990) and Vovk (1993). In order to deal with the conventional normal-theory tests, will need to adapt the ideas of Freedman and Lane (1983a,b) and Beaton (1981).

2.1. Evaluating Melinda's Performance. A crude way of scoring the performance of a person who makes categorical predictions is simply to count how often she is right. But even this crude score will be meaningful only in relation to some baseline. The following example shows how a stochastic story can provide that baseline.

Melinda claims some insight into the behavior of the local train. She claims that at 7:30 she can predict whether the 8:05 train will be on time. As a demonstration, she makes predictions on 100 successive days, and we find that 55 of her predictions are correct. What does this meager success tell us? Does it provide any evidence that Melinda knows what she is talking about?

It appears that Melinda does not know what she is talking about, because she is right barely half the time. Why is being right only half the time so unimpressive? Because we could do as well spinning a coin. Suppose Mary, who knows nothing about the train's behavior, predicts whether it will be on time by spinning a fair coin. In spite of her ignorance, Mary can expect to be right about half the time, too. In fact, Mary has a probability of about one-sixth of being right 55 or more times out of 100.

The simplicity of this example allows us to see clearly that the stochastic story is serving only as a standard for comparison. We compare Melinda's performance to the story, but the story is not about Melinda. I could tell a stochastic story about Melinda if I wanted. I might tell one of these stories:

(1) I might claim that Melinda's own knowledge about the train is such that she can be described as a spectator in a stochastic story. For example, she might know that the train is on time about half the time, without being able at all to predict which half.

(2) I might claim that my knowledge about Melinda's behavior is such that I can be described as a spectator in a stochastic story. For example, I might know that Melinda will predict correctly about half the time, without being able at all to predict which half.

But there is no basis for these stories in what I have told you about Melinda and the train. There is no basis for the first story, because I told you nothing about how often the train is on time. (I only said that Melinda predicted correctly 55 times out of 100. This is consistent with the train always, never, or sometimes being on time.) There is no basis for the second story, because I told you nothing about my knowledge. (Perhaps by 7:29 I always know what Melinda is going to predict and whether she is going to be right.)

The stochastic story is only a thought experiment. The force of the comparison depends, however, on the fact that we could implement the thought experiment if we wished. We are unimpressed by Melinda because we really could predict equally well by spinning a coin (or by using computer-generated random numbers).

A statistical test always involves some score or “test statistic.” Melinda's score is the number of times she predicts correctly. This is an obvious way to score her performance, quite independently of our invention of a stochastic story. The purpose of the stochastic story is to calibrate the score. How large does Melinda's score, say t , have to be in order to provide evidence that Melinda has some insight? We answer this question by calculating, for various values of t , the

probability that Mary's score, say T , will be at least as large as t . Table 1 gives $P(T \geq t)$ for a few values of t . As this table indicates, Mary has a reasonable chance of scoring as well as 55, but she is quite unlikely to score as well as 65. Had Melinda predicted correctly 65 times or more, we would have said that she did better than Mary could reasonably expect to do, and that her performance therefore provides some evidence that she knows more than Mary. (This would say nothing, of course, about the nature of Melinda's knowledge. She may have a way of identifying days on which the train will have difficulties, or she may know that the train is late about 65% of the time and take advantage of this knowledge by always predicting that it will be late.)

t	55	60	65	70
$P(T \geq t)$.16	.03	.002	.00005

Table 1. P-values from spinning a fair coin 100 times.⁶

Of course, our imaginary Mary is only one example of a person who knows nothing about the train. Perhaps someone else who knows nothing about the train could find a more effective way of predicting its performance than flipping a coin. So even if Melinda does better than we could hope for Mary to do, the comparison with Mary is only an argument for Melinda having some insight or knowledge. The argument is a strong one, however. We have had much experience with stochastic stories as standards for comparison, and we do not expect to find a person who is totally ignorant about the train and yet knows how to predict better than Mary.

The probability $P(T \geq t)$, where t is the value of the score actually recorded, is called the “P-value” in the usual accounts of statistical testing. When the P-value

is small (less than the conventional values 5% or 1%, say), the observed score t is called significant—i.e., significantly better than could be expected by chance. I have used P-values in this example in a familiar-looking way, but I have not explained them in the usual way. The usual explanation, which is due to R.A. Fisher, talks about “rejecting a null hypothesis.” The null hypothesis asserts that the data (and hence the score t) was produced by chance, in accordance with a particular objectivistic probability model. We are supposed to reject the null hypothesis when t is large and $P(T \geq t)$ is therefore small, on the grounds that it is easier to disbelieve the hypothesis than to believe that the event $T \geq t$, which actually happened, is so unlikely.

Subjectivists often criticize Fisher's logic on the grounds that it does not justify attention to the event $T \geq t$.⁷ What actually happened in Fisher's story was $T=t$. If we want to claim that the null hypothesis makes what actually happened too surprising, the critics say, we should look at the probability of $T=t$, without amalgamating it with $T>t$, which did not happen. This is not a criticism of my logic. In my story, $T=55$ does not happen (T is Mary's score; 55 is Melinda's score), and attention to the event $T \geq 55$ is justified even before the stochastic story is invented. I observe Melinda's score of 55. I ask myself whether someone who knows nothing about the train can hope to do as well—i.e., can hope for a score T such that $T \geq 55$. I invent the stochastic story precisely in order to study the chances of $T \geq 55$ for one person (Mary) who knows nothing about the train.

In order to put Melinda into Fisher's objectivistic framework, we would have to tell an objectivistic stochastic story about her predictions: they are independent and each is correct with constant probability p . We then test the null hypothesis $p = \frac{1}{2}$, which seems to correspond to Melinda having no real ability to predict. (This null hypothesis is an objectivistic version of the second of the two stochastic

stories about Melinda that I listed earlier.) If Melinda gets 65 predictions out of 100 right, we can reject this story; if she gets only 55, we cannot. The difficulty with this talk, of course, is that the objectivistic story is so ungrounded. Who told us that Melinda's ability to predict is constant from day to day? Why should we accept inferences that seem to depend on such an assumption?

Though the simplicity of this example is atypical of the practice of statistical testing, the lack of grounding for the objectivistic story is quite typical. Statisticians often excuse this lack of grounding by drawing an analogy to the shortcomings of scientific theories, which can be useful even if they simplify reality and remain unconfirmed in many respects. Perhaps our stochastic story about Melinda is a simplification of a more adequate stochastic story, and perhaps analysis of this more adequate story would give the same results. But subjectivists, who tend to doubt the meaningfulness of even the simplest of these objectivistic stories, are not comforted by the thought of making them more complex.

I believe that statisticians take simple statistical tests seriously not because they take the corresponding objectivistic stories seriously as representations of reality, but because they see these stories as standards for comparison. The account of testing I am giving here makes this explicit. This account has not been articulated clearly in the past primarily because the unified informal story of probability, which it uses in an essential way (the spectator must be in the story so that we can compare Melinda with her), has lacked respectability.

Notice that I am arguing for a subjectivist interpretation of the standard test, not for a Bayesian replacement. Bayesian analyses of testing, while vaunting their emphasis on subjectivity, usually take the Fisherian objectivistic story as their starting point. Like the Fisherian analysis, they assume that an objectivistic model generates the data by chance, without reference any observer. We enter as

observers only after this objectivistic model has done its job, and we remain outside the model; our job is to decide whether to believe it.

Before leaving the example, we should note the comparison of Melinda with Mary does not touch on the question of whether the future will be like the past. If Melinda's performance gives evidence that she knew something that helped her predict during the past 100 days, then we may wish to infer that she will continue to know something and continue to make effective predictions during the next 100 days. But this inference goes beyond what we have learned by comparing Melinda with a stochastic story. Neither the story about Melinda nor the stochastic story made any assumption about the 100 days we observed being like other days in the future or the past. In particular, we did not assume that these 100 days were drawn at random from a larger population of days.

2.2. Evaluating a Treatment. It is short step from Melinda to examples that appear in statistics textbooks.

Amanda, who wants to add a new razor blade to her line of toiletries, is trying to decide which of two types of razor blade will be most popular among users, type A or type B. She gives 100 users one blade of each type, and asks them to report back which they prefer. When they do so, 65 report that they prefer type A. Is this strong evidence in favor of type A?

We can deal with this example just as we dealt with Melinda. Melinda made 100 binary predictions. Here, too, we have 100 binary predictions. We can think of the labels on each pair of blades as Amanda's prediction that the blade labelled "type A" will be preferred. Then we can ask whether the success (albeit limited) of these predictions indicates some genuine insight about the superiority of A. In order to rate Amanda's performance, we compare her with Anna, who cannot tell the two

types of blades apart and predicts which blade in each pair will be preferred by spinning a fair coin. Anna, we know, can scarcely hope to do as well as Amanda has done. According to Table 1, the probability she will do as well is 0.002. So Amanda's knowledge must be helping her predict. In other words, something about type A blades makes them more widely preferred.

We may be giving Amanda an unfair advantage in this comparison. We are talking as if Amanda thought type A was better and organized the study to prove the point. If this is so, then the comparison with Anna is fair. But another possibility is that Amanda was uncertain which, if either, of the blades was better, and that she was simply trying to find out. In this case, Amanda has an unfair advantage over Anna. For a fair comparison, we should compare Amanda with Amy, who spins a fair coin in order to label the blades in each pair "A" and "B" and then waits to see how the 100 people's preferences turn out before deciding whether her prediction was that As would be preferred or that Bs would be preferred. Amy's chance of doing as well as Amanda is twice Anna's, or 0.004. (The comparison with Anna is a "one-sided test," while the comparison with Amy is a "two-sided test.")

The objectivistic treatment of this example follows the same path as the objectivistic treatment of Melinda. We posit that each of the 100 people has the same probability p of preferring A over B, and that the preference of each person is independent of the preference of the others. Then we test the null hypothesis that $p = \frac{1}{2}$. Is this probability model any better grounded, any more plausible, or any more meaningful here than in the story about Melinda? I think not.

The comparison of the razor blades with Anna is more complicated than the comparison of Melinda with Mary, because it involves an additional step. First we relate the merit of the razor blades to Amanda's ability to predict, and then we

compare Amanda's ability with Anna's or Amy's. But otherwise the issues are the same. The comparison of Amanda with Anna again makes explicit the real role of the stochastic story; it is really only serving as a standard for comparison.

Here, as in the case of Melinda, we have not touched on whether the future will be like the past. We want, of course, to take the next step and conclude that the majority of future customers will prefer blade A. But our argument based on the comparison with Anna or Amy has no bearing on this next step. Had the 100 people testing the blades been chosen at random from the population of potential future customers, probability arguments might help us make the step into the future, but that is another story.

2.3. Evaluating Lucinda's Ability to Discriminate. Our rating of Melinda, though instructive, was rather crude. We compared Melinda to Mary, who knew absolutely nothing about the train. Mary is easy to beat; if Melinda knows the train is usually late, then she can beat Mary simply by always predicting it will be late. Let us turn, therefore, to a more subtle question about Melinda's performance. For clarity, we will discuss this question for a different forecaster, named Lucinda.

Lucinda claims that by 7:30 she can tell (though she sometimes makes mistakes) whether the 8:05 will be late or not. As a demonstration, she makes predictions on 100 successive days. As it turns out, she predicts 60 times that the train will be late, and she predicts 40 times that it will be on time. We find that 70 of her 100 predictions are right. She was right 55 of the 60 times she said the train would be late, and she was right 15 of the 40 times she said it would be on time. Does this performance provide evidence that Lucinda can tell days the train will be late from days it will be on time?

Table 2 displays the joint performance of Lucinda and the train. The train was late 80 times. Lucinda was right only 70 times, so she could have scored better overall by always predicting the train would be late. But her performance does seem to provide evidence that she can tell a difference between days. The train was late 91.7% of the times she said it would be late (55 out of 60) and only 62.5% of the times (25 out of 40) she said it would be on time.

	Lucinda says train will be late.	Lucinda says train will be on time.	Total
Train is late.	55	25	80
Train is on time	5	15	20
Total	60	40	100

Table 2. Lucinda and the train.

How might we score Lucinda's performance in distinguishing between days? I just suggested one reasonable score: how much more often the train is late when Lucinda says it will be. This is

$$\frac{55}{60} - \frac{25}{40} \approx 0.29. \quad (4)$$

Alternatively, we might measure how much more often Lucinda says the train will be late when it is; this is

$$\frac{55}{80} - \frac{5}{20} \approx 0.44. \quad (5)$$

There are many other possibilities as well; any “measure of association” for the 2x2 table would do. But to interpret any such score we need some kind of baseline or calibration. We need a stochastic story.

Here is a stochastic story that will do. Suppose Lois, who knows nothing about the train, is told that it was late 80 of the last 100 days (this information is

not going to help her, but it helps set the stage). And suppose she is asked to guess 60 of these days. Lacking any other information, Lois uses random numbers produced by her personal computer to choose 60 out of the 100 days, with all the possible choices being equally likely. What are the chances, under these circumstances, for Lois to do as well or better than Lucinda did? In other words, what are the chances that the 60 days she chooses will include 55 or more days on which the train is late? The answer, as it turns out, is about 0.0002. Lucinda has done much better in identifying days on which the train will be late than we could expect from someone who has no knowledge that would help her discriminate.

Notice that the stochastic story has simplified the scoring. Instead of using (4) or (5), we score Lois simply by the number, say T , of her 60 guesses that turn out right. Thus our P-value is $P(T \geq 55)$. We would get the same P-value using (4) or (5), since $T \geq 55$ is equivalent to

$$\frac{T}{60} - \frac{80-T}{40} \geq \frac{55}{60} - \frac{25}{40} \quad \text{or} \quad \frac{T}{80} - \frac{60-T}{20} \geq \frac{55}{80} - \frac{5}{20}.$$

Almost any other measure of association in the 2×2 table will also give the same P-value; since the row and column totals of Lois's table are the same as Lucinda's, Lois can do better only by making T greater than 55.

This example illustrates how comparison with a stochastic story can be effective even though we make arbitrary choices in setting the story up. In order to make Lois's performance comparable to Lucinda's, we asked Lois to guess exactly 60 days. This did not weaken the force of the comparison, because it did nothing to put Lois at a disadvantage relative to Lucinda.

How is the P-value of 0.0002 computed? Readers familiar with combinatorial probability will see that the probabilities for T are hypergeometric;

$$P(T=x) = \frac{\binom{80}{x} \binom{20}{60-x}}{\binom{100}{60}}. \quad (6)$$

We can find $P(T \geq 55)$ by adding these probabilities as x goes from 55 to 60. An approximation using the chi-squared distribution is also available.⁸

Let us now consider the textbook approach to testing Lucinda's performance. There are a number of ways we might proceed, all involving different objectivistic stochastic stories. We might model the behavior of the train, so that we can test whether it behaves differently on days Lucinda thinks are different. We might model the behavior of Lucinda, so that we can test whether she predicts differently on days that are different for the train. Or we might model both together.

Here is a way to model the train. Let X_1 be the number of times the train is late out of the 60 times Lucinda says it will be late, and let X_2 be the number of times it is late out of the 40 times she says it will be on time. Assume that the train is late with probability p_1 on the days Lucinda says it will be late, that it is late with probability p_2 on the days she says it will not be, and that whether it is late on a given day is independent of its performance on preceding days. Under these assumptions, X_1 and X_2 are independent binomial random variables; X_1 has parameters 60 and p_1 , and X_2 has parameters 40 and p_2 . The question whether the days are different has become the question whether $p_1 \neq p_2$. We test the null hypothesis of no difference: $p_1 = p_2$. Under this hypothesis, X_1 and X_2 are independent binomials with a common parameter $p = p_1 = p_2$. As our test statistic, we take the difference

$$\frac{X_1}{60} - \frac{X_2}{40}; \quad (7)$$

this is the score (4) we considered earlier. It turns out that the probability that it will equal or exceed the value we observed for Lucinda, 0.29, is approximately 0.0002, the same as the P-value we obtained by comparing Lucinda with Lois.⁹

Instead of computing the probability of (7) exceeding its observed value unconditionally, it may be better, according to Fisher,¹⁰ to compute its probability

of doing so conditionally, given the observed marginal totals in Table 2. The resulting test is called Fisher's exact test. It is a better test, according to Fisher, because it brings the population of potential repetitions with which we are comparing the actual result closer to that result, and also because it simplifies the analysis. In the unconditional model, the choice of the statistic (7) is somewhat arbitrary, but, as we noted earlier, once the margins of the table are fixed, all measures of association are essentially equivalent. Moreover, the computation of the P-value is simplified. In fact, the conditional probabilities are precisely the hypergeometric probabilities given in (6); Fisher's exact test comes out exactly the same as our comparison of Lucinda with Lois.

It would delay us too long to explore here the other objectivistic models that I have mentioned; suffice it to say that they give similar results and also reduce to Fisher's exact test conditionally. We should also note that there is yet another justification for Fisher's exact test for the 2x2 table in situations where an experimenter is able distribute units over one of the classifications (over the rows or over the columns) of the table randomly. Fisher preferred this justification, but it is obviously inapplicable to Lucinda.

What should we say about the objectivistic approach? Does it make sense? Here, as in the case of Melinda, the objections are obvious. Who told us that the behavior of the train is stochastic? That the probability is the same on every day that Lucinda says the train will be on time? That its behavior on one day is independent of its behavior on another? There are no grounds for these assumptions. Surely a stochastic story can only be justified here as a standard for comparison.

Taking the stochastic story as a standard for comparison allows us make sense of Fisher's intuitions about conditionality, intuitions that many of his

objectivistic successors have found puzzling. Once we acknowledge that Lois is only a standard for comparison, it becomes entirely reasonable that we should design Lois's task so as to maximize its comparability with Lucinda's accomplishment. The vagueness of this desideratum is not a problem, for the desideratum merely serves to help us construct an argument. It does not pretend to exclude any other argument or counterargument.

2.4. The Berkeley Graduate Admissions Data. Table 3 shows the number of men and women who applied for admission to graduate study at the University of California at Berkeley for the fall of 1973, together with the number of each sex who were and were not admitted. These data were first published by Peter J. Bickel and colleagues in 1975.¹¹ These authors were concerned not only with discrimination against women but also with the shortcomings of the objectivistic models used to analyze such questions.

	Admitted	Not admitted	Total	% admitted
Men	3738	4704	8442	44.3%
Women	1494	2827	4321	34.6%
Total	5232	7531	12763	41.0%

Table 3. Graduate admissions at Berkeley in 1973.

As Table 3 indicates, the rate of admission was substantially lower—almost 10 percentage points—for women than for men. Fisher's exact test produces a vanishingly small P-value.¹² The lower rate of admission for women is significant both substantively (10 percentage points is a lot) and statistically (the P-value is practically zero).

Here, as in the case of Lucinda, we can explain the statistical significance in terms of a comparison with Lois. Suppose we tell Lois that 8,442 of the 12,763 applicants are men. We then give her ID numbers for the 12,763 applicants, and we ask her to try to pick out from them 5,232 numbers that identify men. Since she has no way of knowing which of the numbers identify men, she uses her personal computer to choose 5,232 of the 12,763 numbers at random. What is the chance that she will choose as many men as the admissions committees did? This question is answered by Fisher's exact test: the chance is vanishingly small. So the Berkeley admissions process did much better at picking out men than we could possibly expect Lois to do. It picked out more men than could possibly happen by chance.

Though having Lois try to pick out men makes the comparison between Lois and the admissions process simple and rhetorically effective, other ways of setting up the comparison are equally valid and lead to the same conclusion. Suppose, for example, that we ask Lois to pick out 8,442 numbers, trying to include as many admittees as possible. Since she knows nothing about which of the 12,763 numbers represent admittees, she will again choose the 8,442 numbers at random. Suppose Amanda knows which numbers identify men and chooses them. Amanda will have 3,738 admittees among her 8,442 choices, and Lois has practically no chance of doing as well. In fact, her chance of doing as well is given once again by the P-value from Fisher's exact test. So we can conclude that being male predicts admission better than could possibly happen by chance.

An objectivistic treatment of Table 3 would follow the same lines as the objectivistic treatment I sketched for Lucinda and the train. We assume that there is a constant probability of admission for men and a constant probability of admission for women, and we test for the equality of the two probabilities. Or

alternatively, we assume that there is a constant probability of an admittee being a woman and a constant probability of a non-admittee being a woman, and we test for the equality of these two probabilities. As Bickel and his co-authors and many other commentators have pointed out, none these objectivistic assumptions are plausible. As Freedman and Lane (1983b, p. 192) put it, they are known from the first “to be inadequate to describe any aspect of the physical process that generated the data.”

Though the objectivistic models are useless for this example, the comparison with Lois is meaningful. It tells us that something is going on that favors men. This something may or may not be stochastic.¹³ But since it has a stronger effect than could happen by chance, we can reasonably hope that further investigation will yield some insights. Bickel and his colleagues, upon undertaking such an investigation, found that the bias in favor of men was related to how the numbers of places and numbers of men and women applicants were distributed over departments. The rate of admission (number of places available per applicant) was smaller in departments where the proportion of women among applicants was higher. So we can ask why proportionally fewer places were provided in departments to which women more often applied. As it turns out, departments where proportionally more places were provided required, on the average, more mathematical preparation of their applicants. Perhaps society needed a greater fraction of those who were prepared, willing, and asking to study in these demanding fields. This can be contested, but if it is accepted, then the question of why women were being discriminated against in graduate admissions comes down to the question of why they were underrepresented among those prepared, willing, and asking to study in departments requiring more mathematical preparation.

Notes

¹ See especially Shafer 1990a, which describes the informal picture, and Shafer 1992a, which sketches one axiomatization. In both these essays, I used the phrase “ideal picture of probability” for what I am here calling the “informal story of probability.” Unfortunately, the adjective “ideal” seems to have been a source of misunderstanding. One misunderstanding is that the informal story is a representation, less the rough edges, of some reality. This is not my meaning; my theme is that the informal story has many uses; its use to represent a reality we want to understand is only one of these uses. A related misunderstanding is that the merit of the informal story lies entirely in its lack of rough edges—the more ideal the better. This provides an excuse for pushing on to one of the narrower stories, where either the subjective or objective aspects of probability are idealized away.

² Huygens drew an event tree in a manuscript dated 1676 (Edwards 1987, p. 146).

³ Figure 1 calls for the coins to be spun rather than flipped, so that a biased coin—one that is heavier on one side than the other—can exhibit its bias by falling more often on its heavier side. Such a coin is equally likely to fall on either side when it is fairly flipped (Engel 1992).

⁴ For the subjectivist critique, see Berger and Delampady 1987 and the references therein. For a survey of other critiques, see Morrison and Henkel 1970.

⁵ As Stephen Brush (1988) has noted, scientists often use the word “prediction” without regard to whether what is being predicted is already known. In many cases, at least, the credit that a scientific theory earns by predicting an effect does not seem to depend on whether the effect was known before the prediction was made.

⁶ These numbers can be obtained from the normal approximation to the binomial in the usual way: $P(T \geq t)$ is the probability that a normal deviate with mean 50 and standard deviation 5 exceeds $t - \frac{1}{2}$.

⁷ This criticism seems to go back to Harold Jeffreys. See Berger and Delampady 1987, pages 329 and 348.

⁸ See Miller 1986, pp. 47-48.

⁹ The P-value for (6) is usually computed using a normal approximation. Under the null hypothesis, (6) is approximately normally distributed with mean zero and variance $p(1-p)(\frac{1}{60} + \frac{1}{40})$. Since we can estimate p by $\frac{X_1+X_2}{100}$, this implies that

$$\frac{\frac{X_1}{60} - \frac{X_2}{40}}{\sqrt{\frac{X_1+X_2}{100}(1-\frac{X_1+X_2}{100})(\frac{1}{60} + \frac{1}{40})}} \quad (8)$$

should be approximately standard normal. Substituting 55 for X_1 and 25 for X_2 , we find that (8) is approximately equal to 3.6. The probability of a standard normal deviate exceeding 3.6 is approximately 0.0002. As it turns out the square of (8) is equal to the chi-squared statistic used to approximate the sum of hypergeometric probabilities in our comparison of Lucinda with Lois. So the agreement between the two P-values does not depend on the particular numbers we have used in the example.

¹⁰ See Fisher 1973, pp. 89-92.

¹¹ Their article was originally published in *Science* (Bickel et al. 1975). It was reprinted, together with comments by William H. Kruskal and Peter J. Bickel, in Fairley and Mosteller (1977). The issues raised by the data were also discussed by Freedman and Lane (1983b) and Freedman, Pisani, and Purves (1978, pp. 12-15). Inspired by this example, Freedman and Lane (1983b) propose a general way of understanding tests of independence in two-way contingency tables. My discussion here is influenced by their proposal but does not follow it. The comparison I suggest with a unified stochastic story is, I think, better motivated and more persuasive than Freedman and Lane's purely "descriptive" and "non-stochastic" treatment, and it applies only to 2x2 tables.

¹² The chi-squared statistic, which has one degree of freedom, is 110.8.

¹³ On our unified understanding of stochasticity, it surely was not stochastic at the beginning of the investigation by Bickel and his colleagues, for stochasticity requires an observer, and no one had been closely observing what was going on.

References

- Berger, James O., and Mohan Delampady (1987). Testing precise hypotheses (with discussion). *Statistical Science* **2** 317-352.
- Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187** 398-404.
- Brush, Stephen G. (1988). Prediction and theory evaluation. *Science* **246** 1124-1129.
- Dawid, A.P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association* **77** 605-613.
- Dawid, A.P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A* **147** 278-292.
- Dawid, A.P. (1985). Calibration-based empirical probability. *The Annals of Statistics* **4** 1251-1273.
- Dawid, A.P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences* (eds S. Kotz, N.L. Johnson, and C.B. Read), vol. 7, pp. 210-218. New York: Wiley-Interscience.
- Dawid, A.P. (1990). Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society, Series B* **53** 79-109.
- Edwards, A.W.F. (1987). *Pascal's Arithmetic Triangle*. Oxford University Press, New York.
- Engel, Eduardo M.R.A. (1992). *A Road to Randomness in Physical Systems*. Lecture Notes in Statistics #71. New York: Springer-Verlag.
- Fairley, William B., and Frederick Mosteller (1977). *Statistics and Public Policy*. Reading, Massachusetts: Addison-Wesley.

- Fisher, R.A. (1973). *Statistical Methods and Scientific Inference*. Third Edition. New York: Macmillan.
- Freedman, David, and David Lane (1983a). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics* **1** 292-298.
- Freedman, David, and David Lane (1983b). Significance testing in a nonstochastic setting. Pp. 185-208 of *Lehmann Festschrift* (Peter Bickel et al., eds), Wadsworth.
- Freedman, David, Robert Pisani, Roger Purves, and Ani Adhikari (1991). *Statistics*. Second Edition. New York: Norton.
- Miller, Rupert G., Jr. (1986). *Beyond ANOVA, Basics of Applied Statistics*. Wiley.
- Morrison, Denton E., and Ramon E. Henkel, eds. (1970). *The Significance Test Controversy*. Chicago: Aldine.
- Shafer, Glenn (1985). Conditional probability (with discussion). *International Statistical Review* **53** 261-277.
- Shafer, Glenn (1990a). The unity and diversity of probability (with discussion). *Statistical Science* **5** 435-462.
- Shafer, Glenn (1990b). Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning* **4** 323-362.
- Shafer, Glenn (1992a). Can the Various Meanings of Probability be Reconciled? Pp. 165-196 of *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, ed. G. Keren and C. Lewis, Hillsdale, N.J.: Lawrence Erlbaum.
- Vovk, V.G. (1993). The logic of probability. To appear in the *Journal of the Royal Statistical Society, Series B*, **55** No. 2.