Is Everything Stochastic?[1]

Glenn Shafer[2]

Prisme N° 20

December 2010

[1] This text is based on the transcription of the talk that Glenn Shafer gave at the Cournot Centre's "Probabilism Sessions" seminar held on 13 October 2010 in Paris. The video is available at www.centre-cournot.org.

[2] Glenn Shafer is Board of Governors Professor at Rutgers University. He serves as faculty director of the doctoral program of the Rutgers Business School – Newark and New Brunswick. He is also a professor at the Computer Learning Centre, Royal Holloway College, University of London. Shafer is the author of five books and numerous research papers that have appeared in journals of statistics, philosophy, history, psychology, computer science, economics, engineering, accounting and law.

Summary

Kolmogorov said no, Popper said yes. My sympathies lie with Kolmogorov, the old-fashioned empiricist.

In the on-line setting, where we see previous outcomes before making the next probability forecast, we can give probabilities that have objective value because they pass statistical tests. This accounts for the success of many adaptive methods, and it is related to Leonid Levin's notion of universal probability distributions. It tells us that yes, everything is stochastic, but in a sense that is empirically empty insofar as it is not falsifiable.

When we understand that success of adaptive methods does not depend on the world being stochastic in a falsifiable sense, we may want to be more parsimonious in causal modeling and more open to non-standard methods of probability judgement.

Contents

1. Introduction

The question "Is everything stochastic?" can be interpreted in different ways. I will begin by discussing three interpretations: Andrei Kolmogorov's, Karl Popper's, and a new, slightly mischievous interpretation involving a game in which probabilities serve as predictions and betting rates.

The game is one of *on-line prediction*: new predictions are made only after previously predicted outcomes are observed. In this on-line setting, it turns out, a forecaster can give probabilities that are objective inasmuch as they pass statistical tests. I will explore implications of this result for the existence, meaning, and use of probabilities.

2. Three ways of understanding the question

Is everything stochastic? Does every event have an objective probability? More precisely: In given conditions, once you've fixed the conditions or the experimental arrangement, does every outcome have an objective probability?[3] Kolmogorov and Popper, both writing in the 1950s, disagreed on this. Kolmogorov said no. Popper said yes. I will also say yes.

I agree with Kolmogorov given the way he understood the question. I do not think so highly of the way Popper understood and answered the question. But I will interpret the question in a different way than either Kolmogorov or Popper did, and so I will be able to give the same answer as Popper while not agreeing with him.

Kolmogorov

Here is what Kolmogorov said in his article on probability in the *Great Soviet Encyclopaedia* in 1951:

[3] Some readers may prefer to use the word "stochastic" in various other ways. I ask them to pardon my using it to get their attention and to continue reading if they are interested in the question posed here: whether every outcome has an objective probability once you fix an experimental arrangement.

In many contexts, "stochastic" and "deterministic" are thought of as opposites. But here I am treating "deterministic" as a special case of "stochastic". When an event is sure to happen, it has an objective probability, namely one, and is therefore stochastic in the sense in which I am using the word.

> *Not every event has a definite probability. The assumption that a definite probability in fact exists for a given event under given conditions is a hypothesis that must be verified or justified in each individual case.*[4]

He was thinking: "OK, there is an experiment you can do over and over; sometimes you will get heads; sometimes you will get tails. Will the sequence of heads and tails behave the way the mathematical theory of probability predicts?" As Kolmogorov realized, there are many ways the sequence might fail to behave like a probabilistic sequence. The fraction of heads might fail to converge to a limit. Even if it does converge, it might do so in an inappropriate way. It might converge from above, for example. Whether it converges and in the right way is something you've got to test and check empirically. That was his viewpoint, and it seems perfectly sensible.

At the time when he formed his views on probability – in the 1920s and 1930s – Kolmogorov's view that objective probabilities only sometimes exist was not particularly original or particularly unusual. It was by far the majority view among mathematicians and economists and philosophers who thought about probability. After the most recent economic debacle involving bankers who believed in objective probabilities and mathematicians who were supposed to be able to calculate their values, some commentators have remembered that Frank Knight and John Maynard Keynes, like Kolmogorov, had mentioned that they didn't think everything has a probability. An odd heretical opinion, these commentators suggested, but maybe worth thinking about for a second. I think we should dig a little deeper into the history of twentieth-century thinking about probability and recognize that Knight and Keynes and Kolmogorov were saying what almost *everybody* thought in the 1920s and 1930s. You can find Irving Fisher saying it too. *No one* thought everything had an objective probability. Bruno de Finetti thought everything had a probability, but he was talking about subjective probability. In his view, nothing had an objective probability. Emile Borel wanted to compromise: you can always give a probability, and sometimes it is more objective than other times. But from 1900 through to the 1930s, everyone would have agreed: "Of course, objective probability is something

[4] This is my abridgement of a passage from the English version of the encyclopedia.

very special; not every event has an objective probability; that's a very special hypothesis."

Kolmogorov is remembered for his mathematics, not his philosophy. Among his many mathematical achievements is a small monograph that he published in 1933, *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Systematizing ideas that others had been developing for decades, the monograph proclaimed that mathematical probability is a branch of measure theory. After World War II, mathematicians came to regard Kolmogorov's measure-theoretic axioms as the definitive foundation for probability. Philosophers and statisticians had no choice but to accept their judgement on this point, but it posed puzzles. If measure theory is the right mathematical framework for probability, what does this tell us about the meaning of probability, about its interpretation, about its use to describe the world, about what probabilities really are?

Popper

Karl Popper is known for the thesis that science progresses by falsifying hypotheses, not by verifying them. This thesis became known around the world as a result of the great success of his *Logik der Forschung,* which he published in Vienna at the end of 1934. When he wrote *Logik der Forschung*, Popper knew nothing about measure theory. He'd never heard of Kolmogorov. In 1938, he published his own axioms for probability, with no reference to Kolmogorov. It was only when he returned to England from New Zealand in the late 1940s, after concentrating on political philosophy for some years, that he found out that everybody now thought that probability had been axiomatized by Kolmogorov. Kolmogorov was now the king. So Popper was faced with the problem of how to interpret the new measure-theoretic framework.

Popper decided to interpret measure-theoretic probabilities as propensities. Here's the way he says it in his *Realism and the Aim of Science*:

> *I suggest* a new physical hypothesis*: every experimental arrangement generates propensities which can sometimes be tested by frequencies.*[5]

It seems that whether we can test the propensities is not important for Popper. They are there anyway.

Realism and the Aim of Science was the first of three books that together constituted Popper's *Postscript to the Logic of Scientific Discovery*. In the second, *The Open Universe*, he makes room for propensities by rejecting all varieties of determinism – scientific, religious, and metaphysical, as he calls them. In no sense, he insists, is there always a rule that would tell us, if we knew the rule and all the information about the current state of the world that the rule needs as inputs, what will happen next. But he seems to have believed that there is always a rule that gives probabilities for what will happen next. Popper called himself a *metaphysical realist*. You might also call him a *metaphysical stochasticist*.

The contrast with Kolmogorov's thinking is striking. Kolmogorov was thinking about actually repeating an experiment and seeing what happens. If you do not find a stable frequency, converging in the right way, then the event does not have an objective probability. Popper was not so concerned with real repetitions. Maybe you can't repeat the experiment, but anyway, you can *imagine* repetitions, and so there will be a *virtual* frequency (virtual means imaginary, I guess), and that's all you need, even if the imagined repetition is impossible. So there is always an objective probability.

I have singled out Popper for discussion because his way of thinking about probability became popular among statisticians and other users of mathematical probability after World War II. I am not sure that he was influential in this respect,[6] but he was representative.

---

[5] The book was largely written in the 1950s but not published until 1983. The quotation, which I have abridged, is on p. 360.

[6] Popper is probably the philosopher of science most often read by scientists, and his authority surely helped legitimize the idea that probabilities are propensities. But I doubt that his writing about probability was influential among mathematicians. While I admire his discussion of determinism in *The Open Universe*, I find his discussion of probability in *Realism and the Aim of Science* confused and ill-informed.

Before the war, most mathematicians who worked on or used probability (like most other scientists) were empiricists rather than realists.[7] They were too close to the invention of real numbers to think real numbers were real. They were wary of infinity. When Kolmogorov argued for the axiom that tells us that the probability of the disjunction of an infinite number of disjoint events is the sum of their probabilities, he did not argue that this reflects some truth about reality. On the contrary, he argued that the axiom is a mathematical convenience, mathematically useful but harmless in applications because you will never see an infinite number of events. For Kolmogorov and Borel, mathematics could serve science only when there was a way of relating it to the world of experience, which is necessarily finite. The concept of infinity is a wonderful mathematical tool, because it simplifies and thereby clarifies our mathematical pictures. Messy details that persist and become ever more complicated when we consider a larger and larger number of objects often disappear in the limit. But in order to use an infinitary mathematical picture as a practical model, we must find a way to crawl back from infinity and deal with the messiness of the finite.

I don't know whether Popper ever read what Kolmogorov had to say about relating measure-theoretic probability to the world of experience. But it would have made little difference. Popper was concerned not with the world of experience but with reality, and he had no scruples about supposing that reality contains any number of infinities and real numbers. He adopted his realism in self-conscious opposition to the empiricism that the Vienna Circle shared with most scientists in the 1930s. After the war, he was not so lonely. The collapse of the Vienna Circle's version of empiricism, logical positivism, created a large cohort of realists among philosophers. The vast expansion of mathematical training for engineers, statisticians, and economists – training that was often more rigorous than nuanced – produced an army of applied mathematicians who thought of real numbers and infinity as real in a way that Kolmogorov and Borel never could.

---

[7] This is not to say that they would have described themselves as empiricists and denied being realists; they had other ways to describe their views. For a contemporary empiricist's view of the opposition between empiricism and realism, see van Fraassen (1980). For a realist's view, see Psillos (1999).

Stochastic processes

I want to call attention to an aspect of my quotations from Kolmogorov and Popper that now seems outdated. They both considered how an individual probability can be manifested as a frequency in a sequence of independent identically distributed trials. They did not give an equally fundamental role to stochastic processes in which trials are not independent and identically distributed. In the 1930s, independent and identically distributed trials were still relatively central to probability's use in statistics, economics, and other applied fields. But as Jerzy Neyman pointed out in 1960, probability has now moved on. More often than not, our model is now a more complex stochastic process.

The concept of independent and identically distributed trials is probability theory's representation of the idea of repeating an experiment. Suppose *p* is the probability for *E* each time the experiment is repeated. Set

$$y_i = \begin{cases} 1 \text{ if } E \text{ happens on the } i\text{th trial,} \\ 0 \text{ if } E \text{ does not happen on the } i\text{th trial.} \end{cases}$$

Then *E*'s frequency in the first *n* trials is

$$\frac{\sum_{i=1}^{n} y_i}{n}$$

The classical theorems are concerned with the difference

$$\frac{\sum_{i=1}^{n} y_i}{n} - p. \qquad (1)$$

Bernoulli's theorem (*weak law of large numbers*) says that (1) is small with high probability when *n* is large. Borel's theorem (*strong law of large numbers*) says that (1) converges to zero with probability one as *n* tends to infinity. The *law of the iterated logarithm* tells us the rate at which (1) approaches zero as *n* grows, again with probability one.

Kolmogorov and Popper seem to have agreed that these theorems must be satisfied in order for p to be the objective probability for E.[8] They disagreed about whether we can be sure *a priori* that there exists a p for which they are satisfied. According to Kolmogorov, we have to check empirically. According to Popper, an experimental setup defines such a p even if we cannot repeat the experiment well enough to get the behavior we expect or even if we cannot repeat it at all.

This discussion seems outdated because it is now natural to consider more general theorems. In a stochastic process, there may not be any event that has the same probability on every trial, but we can consider different events as we go along. Let E_i be an event settled on the i^{th} trial, and let p_i be the probability given to E_i by the model after the previous trials have been completed. In this case, we replace (1) by

$$\frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} p_i}{n} = \frac{\sum_{i=1}^{n} (y_i - p_i)}{n} \qquad (2)$$

It is now this number that should become very small and even converge to zero with probability one as n grows. Does it? Do the laws of large numbers and law of the iterated logarithm still hold? Yes. We can prove *martingale versions* of these classical theorems. But they are now theorems about a sequence of probabilities $p_1, p_2, \ldots$, each of which may depend on how preceding events came out, not theorems about a single probability p specified in advance. We are no longer talking about a "frequency interpretation" or a "propensity interpretation" of a single probability p.

It is easy to adapt Kolmogorov's attitude to this more general situation: we merely continue to insist that the model needs to be checked empirically. There are many possible tests; we can check the martingale law of large numbers for various sequences of events, and we can also check many other predictions that the model makes with high probability. For the most part, these tests will not involve finding frequencies that match individual probabilities in the model, and they will leave almost all the probabilities in the model with no such frequency interpretation.

[8] In Popper's case, it may be more accurate to say merely that he considered the laws of probability to apply. I am not sure that Popper ever discussed the law of the iterated logarithm, and passages in *Realism and the Aim of Science* suggest that he did not understand the difference between Bernoulli's and Borel's theorems.

It is not so clear how to adapt Popper's attitude. His idea was to imagine what would happen in an infinite number of trials. Would he be content for this imagined infinite sequence of trials to consist of a single run of the infinitely long stochastic process in which (2) converges to zero and other events that are supposed to have probability one also happen? Or would he insist on extending his imagination to interpret all the individual probabilities at every point of time in the model?[9]

As you see, I sympathize with Kolmogorov's empiricism and find Popper's realism unrealistic. My purpose, though, is not to mock Popper. Rather, it is to set the stage for my own way of asking whether everything is stochastic. Following Kolmogorov, I see the question as one about statistical tests. Can we give probabilities that pass them?

Objective probabilities as predictions

Does every event have an objective probability? For generality, let's ask the question in the spirit of stochastic processes. Do not assume that we are repeating the same experiment over and over, whatever "same" means. Only assume that there is a sequence of events, that our event is one of them, that at every step we assign a probability to the next event, and that we want these probabilities to pass statistical tests. For me, passing statistical tests is enough to make probabilities objective. What more can you ask for?[10] That's my point of view. And with that point of view I'm saying yes, *every* event that is in a sequence of events, where we've agreed on the sequence, has an objective probability. We can assign the events probabilities that pass statistical tests.

I'm talking about a stochastic process where you don't start with probabilities. You don't have any theory that tells you probabilities in advance. Well,

[9] Isabelle Drouet (in press) has reviewed the paradoxes that abound when we insist on applying Popper's idea to every probability and conditional probability in a model.

[10] Especially before World War II, many continental mathematicians subscribed to "Cournot's principle", which says that the mathematical theory of probability makes contact with phenomena only by means of its predictions with probability equal or close to one; see Shafer (2007). It follows from this principle that the validity of a probabilistic theory can be tested only by checking predictions it makes with probability equal or close to one. For a review of more recent consideration of these ideas, which seems to have lost contact with the earlier history, see Hennig (2007).

we haven't really even decided yet whether the process is stochastic, so let's just call it a "process". Can I make up probabilities for the process that will pass statistical tests?

If you want me to make up all the probabilities before I see any of the outcomes, if you want me to announce now probabilities for what's going to happen tomorrow, the next day, the day after that, and so on, then I am not sure I can do that. But give me an easier job. Ask me to give a probability for tomorrow, and then tomorrow – after I've seen what actually happened, how things turned out – ask me to give a probability for the next day. Then the next day, when I see how that turns out I give a probability for the day after, and so on and so forth. In other words, let me give a probability for each trial after I have seen the outcomes for previous trials. This is in the spirit of what folks in machine learning now call "on-line" prediction. It turns out that good on-line probability prediction is possible.[11]

This is the technical content of my presentation: I want to persuade you that I can give probabilities sequentially that pass statistical tests. So in this sense, everything is stochastic.

Of course, if everything is stochastic, then maybe stochastic is not what it was cracked up to be. Or maybe you are letting me off too easily by allowing me to see all the previous outcomes. But let's come back to these points later.

3. A forecasting game

We know how to do statistical testing when we are given probabilities. But I am proposing to you that we do it without any probabilities. No probabilities at the outset, anyway. I want to make the probabilities up as we go along. The only way to make mathematical sense of this is to set it up as a game. One player makes up the probabilities. Another player decides on the outcomes. We will also need a player who does the testing.

I will continue to discuss only the simplest case, where the outcome is binary: yes or no, heads or tails, one or zero. We can deal with much more general cases. The outcome space can even change from trial to trial. But let's keep things as simple as possible.

[11] See Cesa-Bianchi and Lugosi (2006).

Forecaster and Reality

Let's start with two players; call them Forecaster and Reality. Forecaster gives a probability; in other words, a number between zero and one. Then Reality announces the outcome: 0 for tails, say, and 1 for heads. They do this over and over forever. We could talk instead about a finite number of trials, but that would be more complicated.

I will assume that the game is one of perfect information. Forecaster makes the first move by announcing his probability p_1. After seeing Forecaster's move, Reality announces the outcome y_1. Then they do it again: p_2 and y_2. And so on. Each player sees the other player's moves as they are made.

In this framework, we can prove theorems about what one or the other player can accomplish.[12] Such a theorem says that one player has a strategy that will achieve a certain goal regardless of how the other player moves. Many of the theorems generalize classical theorems in probability and have them as corollaries. Another theorem, which I am not quite ready to state precisely, will say that Forecaster can give probabilities that will not be refuted by statistical tests.

Often we can relax the assumption that each player sees the other's moves. If one player has a winning strategy, then it obviously remains a winning strategy if his opponent has less information. So it's not always essential for the results that everybody have all the information. But of course we're not going to test Forecaster using information he doesn't get to see. Players may also get other information as the game proceeds. I don't have time to say much about this, but in practical work you usually have some x's to help you predict the y's.

To define the game, you must you say more than how it's played; you must also tell how the winner is determined. I'll get to that. But we have already defined the game well enough to talk about strategies. What is a strategy for Forecaster? It is a rule that tells him what probability to announce depending on how the other player, Reality, has moved so far. So it is more or less the same thing as a probability distribution for Reality's moves $y_1,y_2,\ldots$. If Forecaster begins with a probability distribution P for $y_1,y_2,\ldots$, he can give as p_1 the probability P gives for $y_1=1$, and then after he sees y_1, say $y_1=0$, he can give as p_2 the probability that P gives for

[12] See, for example, the theorems in Shafer and Vovk (2001).

$y_2=1$ conditional on $y_1=0$. And so on. In other words, he always conditions P on what he has seen so far. So in this setting,

probability distribution for Reality = strategy for Forecaster.

The idea of conditioning a probability distribution P on what has been seen so far in order to obtain a probabilistic prediction for what will happen next is often called "Bayesian". Here, however, the "prior" distribution P is not constructed in a traditional Bayesian way. It is not constructed by thinking about someone's beliefs. Nor is it constructed by averaging a statistical model (a class of probability distributions) with respect to subjective beliefs about which probability distribution in the statistical model is right. Instead, as we will see, it is chosen to defeat a (more or less) universal test.

Skeptic

How can you perform a statistical test when you don't start with a probability distribution? Well, the statistical tests are performed by another player, Skeptic, who bets according to the probabilities given by Forecaster. You can think of Forecaster's probability as the price for a ticket that pays Reality's move in dollars: \$1 or \$0. Skeptic can decide how many of those tickets to buy. He starts the game with a capital of one dollar, and he tries to parlay that it into a large or infinite amount. And he tries to do so without risking his current capital ever being negative. If you risk your capital being negative, that means you are risking somebody else's money, because you're making a bet on credit. So the idea is that Skeptic is trying to multiply the capital he risks by a large or infinite factor.

In standard probability theory, the probability is zero that you will multiply the capital you risk by an infinite factor – i.e., that a non-negative random variable will come out infinitely greater than its expected value. The probability is small that it will come out many times as large as its expected value. So Skeptic is trying to do something that has a small or zero probability. The normal way of doing statistical tests is to reject the hypothesis that a probability distribution is right if something to which it gives very small or zero probability happens. Here, where we don't start with a probability distribution, we can still reject the hypothesis that Forecaster is a good

forecaster if Skeptic manages to multiply his capital by a large or infinite factor. The idea of doing statistical testing this way is due to Jean André Ville, who defended his doctoral thesis before a jury headed by Emile Borel in 1939.[13]

I am going to concentrate on the case where Skeptic tries to multiply his capital by an infinite factor, which is like probability zero, and where he has an infinite number of trials in which he can try to do so. The case where Skeptic tries only to multiply his capital by a large finite factor in a finite number of trials is too complicated to present here. For example, I will talk about the strong law of large numbers rather than the weak law of large numbers, because I don't have time to get into the details of how much the capital can be multiplied in how many trials if the frequency is a given distance from the probability, and so on. Please don't think that there is no finitary version of the story; like Kolmogorov, I'm using infinities only for simplicity.

What Skeptic can do

To illustrate what Skeptic can do, I am going to show you Ville's game-theoretic proof of the strong law of large numbers.

For simplicity, let's assume that Forecaster always gives 1/2 as his probability. This is not an assumption about how Reality will choose y_n. I'm not making any stochastic assumption. I'm not assuming that the objective probability for $y_n = 1$ is 1/2 in the sense you're accustomed to. The probability is 1/2 only in the sense that Forecaster authorizes Skeptic to make bets at even odds.

As I've already explained, Forecaster's move, 1/2 in this case, is the price of a ticket that pays Reality's move y_n in dollars, \$1 or \$0. The net payoff from one ticket is $\$(y_n - 1/2)$. Skeptic moves by choosing how many tickets to buy; call his move s_n. It can be any real number, positive or zero or negative. If s_n is positive, Skeptic is betting on $y_n = 1$. If s_n is negative, Skeptic is really selling tickets rather than buying them; he is betting on $y_n = 0$. His total payoff will be the number of tickets times the net payoff for each ticket: $s_n(y_n - 1/2)$.

[13] For more on its relation to other methods of statistical testing, see Shafer et al. (in press) and Dawid et al. (2011).

The law of large numbers says that the fraction of Reality's moves that are 1s should converge to 1/2. Ville proved that Skeptic has a strategy that (1) keeps his capital from ever becoming negative no matter what Reality does, and (2) makes him infinitely rich if Reality does not make the convergence happen. Reality can obviously avoid the convergence, and Reality can obviously keep Skeptic from making money, but Ville's theorem says Reality cannot avoid both. Skeptic has a way of playing so that one or the other happens.

In order to put the whole story into formulas, let us write K_n for Skeptic's capital after the n^{th} trial. We set K_0, his capital at the beginning, equal to 1. Then we can describe the game like this:

$K_0 := 1.$
FOR $n = 1, 2 \ldots$:
  Skeptic announces $s_n \in R$.
  Reality announces $y_n \in \{0, 1\}$.
  $K_n := K_{n-1} + s_n \left(y_n - \frac{1}{2}\right).$
Skeptic wins if both
  (1) K_n is never negative, and
  (2) either $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{2}$ or $\lim_{n\to\infty} K_n = \infty$.

Notice that there are now only two players, Skeptic and Reality. (Because we have fixed Forecaster's strategy, there is no need to list him as a player.) One of the two players will win; the other will lose.[14]

Ville's theorem says that Skeptic has a winning strategy. Ville constructed such a strategy explicitly, giving a formula for Skeptic's move s_n as a function of Reality's previous moves. The formula says

[14] In games between two players in which one wins and the other loses, it is conventional to say that the winner gets 1 and the loser gets -1; we then say that the game is *zero-sum*. This score of 1 or -1, not the capital K_n, is the "utility" that Skeptic is trying to maximize. We are not using game theory as it is usually used in the social sciences, where each player tries to maximize a utility function that aggregates complex wants and preferences, and so has many possible values.

$$s_n(y_1, \dots, y_{n-1}) = \frac{4\left(r_{n-1} - \frac{n-1}{2}\right)}{n+1} K_{n-1}, \text{ where } r_{n-1} = \sum_{i=1}^{n-1} y_i.$$

This formula tells Skeptic to bet a certain fraction of his current capital on heads, that is, $y_n = 1$. The fraction depends on how much the current number of heads differs from half of the number of trials so far. If there are too many heads, Skeptic bets that Reality will produce more heads, and to defeat Skeptic, Reality must move back towards one half, and *vice versa*.

It is easily verified by induction that this strategy produces the capital

$$K_n = 2^n \frac{r_n!\,(n - r_n)!}{(n+1)!}.$$

Expanding the factorials by Stirling's formula and applying the Kullback–Leibler inequality, Ville arrived immediately at the game-theoretic law of large numbers: for the capital not to be bounded, r_n/n has to converge to 1/2. Thus, Skeptic gets infinitely rich if Reality doesn't choose heads ($y_i = 1$) half the time.

It is easy to convert this remarkably simple proof of the game-theoretic strong law of large numbers into a proof of the measure-theoretic strong law, which says that the convergence will happen except on a set of measure zero. Just use the fact that the probability of any strategy multiplying the capital it risks by an infinite factor is zero.

Testing mere probabilities

Now let's bring Forecaster back and allow him to announce whatever probability he wants. There are now three players. Forecaster tries to give probabilities that will pass statistical tests. Skeptic implements the tests by trying to get rich betting at Forecaster's probabilities. Reality decides. Here is a complete description of this more general game:

$K_0 := 1.$
FOR $n = 1,2, \ldots$:
Forecaster announces $p_n \in [0, 1]$.
Skeptic announces $s_n \in R$.
Reality announces $y_n \in \{0, 1\}$.
$K_n := K_{n-1} + s_n(y_n - p_n)$.

Skeptic wins if both
(1) K_n is never negative, and
(2) either $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i) = 0$ or $\lim_{n\to\infty} K_n = \infty$.

The proof I just gave for the case when p is always 1/2 generalizes to a proof that Skeptic has a winning strategy in this game. This is Ville's game-theoretic proof of the martingale strong law of large numbers that I discussed in Section 2.

Ville published this result in 1939, but it adds a dimension to statistical testing that is still often overlooked. In Section 2, I recalled that we can statistically test a probability distribution for a stochastic process using the results of a single realized path of the process, even though the successive steps are not independent and identically distributed according to that distribution. Ville's result implies that we do not even need a whole probability distribution for the process. All we need in order to do statistical testing is a sequence of probability forecasts to test. This point was first made crystal clear by Philip Dawid (1984).

In *The Open Universe*, Popper emphasized that science's ability to predict can be very scattered and fragmentary. Ville's game-theoretic testing allows us to push Popper's insight farther than Popper did, for it shows that probabilistic predictions can be legitimately tested even when they are more fragmentary than predictions derived from probability measures. A probability measure is a closed system; it specifies in advance the possibilities for information and the probability forecasts (i.e., conditional probabilities) these different possibilities will produce. But the game laid out above says nothing about when Forecaster will make his next probability prediction or what information he may use.

4. How to make probability forecasts

Having studied what Skeptic can do, let's now think about what Forecaster can do. To get started, let's first think about what Forecaster can do in the worst case, when Skeptic and Reality team up against him. It is obvious that he cannot make good probability predictions in this case, but what is obvious turns out to be wrong, and this will point us to the forecasting method that Vladimir Vovk has dubbed "defensive forecasting".

Hilary Putnam's strategy against Forecaster

As I said, it is obvious that Forecaster doesn't have a chance when Skeptic and Reality team up against him. You see immediately what they will do. Reality will do the opposite of what Forecaster says is likely, and Skeptic, knowing what Reality will do, will bet against what Forecaster says is likely. Skeptic will make steady money this way – an infinite amount if play goes on forever.

Hilary Putnam, the philosopher, once thought it worthwhile to write down in detail exactly what the strategies for Skeptic and Reality would look like. Here is how it goes. Suppose Reality makes the event happen whenever Forecaster's probability is less than or equal to 1/2 and makes if fail whenever Forecaster's probability is more than 1/2.[15] Suppose Skeptic knows Reality will do this and bets accordingly, but limits the stakes to \$1 each time; he wants to beat Forecaster, not humiliate him.

FOR $n = 1, 2, \ldots$
  Forecaster announces $p_n \in [0, 1]$.
  Skeptic announces $s_n \in R$.
  Reality announces $y_n \in \{0, 1\}$.
  Skeptic's profit $= s_n(y_n - p_n)$.

Reality makes Forecaster look as bad as possible:

[15] It doesn't matter what Reality does when Forecaster's probability is exactly 1/2, provided that Skeptic knows what he will do.

$$y_n = \begin{cases} 1 & \text{if } p_n < 0.5 \\ 0 & \text{if } p_n \geq 0.5. \end{cases}$$

Skeptic then makes steady money:

$$s_n = \begin{cases} 1 & \text{if } p_n < 0.5 \\ -1 & \text{if } p_n \geq 0.5. \end{cases}$$

Skeptic's profit $s_n(y_n - p_n)$ is $\pm(\pm 0.5) = 0.5$. He makes 50 cents every time.

This way of beating Forecaster looks very convincing at first, but it has a feature that is unreasonably artificial: Skeptic's testing strategy, which depends only on Forecaster's last move p_n, is discontinuous as a function of p_n. To wit:

$$s_n(p) = \begin{cases} 1 & \text{if } p < 0.5 \\ -1 & \text{if } p \geq 0.5. \end{cases}$$

Why is this unreasonable? Because the notion that we can tell to which side of an exact value a real number lies is an idealization that does not represent distinctions we can actually make. A real-valued function that represents an operation we can actually perform should be continuous, so that a deviation in its input too small to notice will not make a difference in its output great enough to notice.[16] Perhaps we can allow Reality to be discontinuous; Reality will be whatever Reality is; God does as he pleases. But Skeptic represents a test, a test that might be conducted by a human or a computer or at least by some mechanism that leaves a record we can discern. It makes no sense to represent Skeptic as behaving discontinuously.

This way of looking at the matter is supported by my 2001 book with Vovk, *Probability and Finance, It's only a Game*, where we show that conventional statistical tests, including those based on the law of large numbers, can all be implemented by continuous betting strategies. This is hardly surprising; who would want to rely on a statistician whose decisions are based on splitting hairs he cannot see? So it is reasonable to assume that Skeptic will play continuous betting strategies and to limit our demands on Forecaster to asking him to beat such strategies.

[16] L.E.J. Brouwer said it this way: a constructive function of a real variable must be continuous.

If you are too accustomed to using discontinuous functions in applied mathematics to accept this argument, there is an alternative argument that you may find more appealing. It says, OK, if you're going to require Forecaster to beat these imaginary discontinuous strategies, give him a little break, and let him randomize a little. Instead of asking him to announce his probability to infinite precision, ask him only to announce some tiny, tiny interval, and after his opponents have moved, then *you* choose an exact value at random from that tiny, tiny interval. For example, if he is thinking of announcing 1/2, let him instead just announce the interval $1/2 \pm 1/10^{100}$, and then after the other players have moved, you choose at random (if you think you can really do such a thing) an infinitely precise number in this interval. One way or another, give Forecaster a break from the fiction of infinite precision. Either let him randomize a tiny bit or else only ask him to beat continuous strategies. The two ideas are pretty close. A tiny bit of randomization will produce a tiny bit of averaging in the payoff, all you need to make it continuous.

Defensive forecasting

Now I'm going to show you how Forecaster can beat any given continuous strategy for Skeptic, regardless of how Reality moves.

For good measure, let's make the story a little more realistic by allowing Forecaster to use some additional information. Call it x_n. The other players can also be allowed to see x_n; to make this clear we will assume that it is announced before any other moves are made. (Remember that the game is one of perfect information; all the players see whatever is announced.) We let Reality announce it. Who else?

FOR $n = 1, 2 \ldots$
  Reality announces $x_n \in X$.
  Forecaster announces $p_n \in [0, 1]$.
  Skeptic announces $s_n \in R$.
  Reality announces $y_n \in \{0, 1\}$.
  Skeptic's profit $= s_n(y_n - p_n)$.

I am going to prove to you that if Skeptic follows a known strategy that is always continuous in the last move made by Forecaster, then Forecaster can play in such a way that Skeptic will never make any money at all.

One way of formalizing the idea that Forecaster knows the strategy Skeptic is using would be to have Skeptic announce this strategy at the beginning of the game. We could then take Skeptic out of the game and simply use the known strategy to compute his profit on each trial. But we can still prove the theorem if we ask a little less of Skeptic. Instead of asking him to tell his strategy at the outset, ask him only to tell his strategy for the particular move right after reality announces x_n. At this point in the game, Forecaster's p_n is the only information not yet known that Skeptic's strategy is allowed to use in choosing s_n. So in order to tell how he will choose s_n, Skeptic can just give a function S_n on [0,1] that will be applied to p_n to determine s_n. This is the function that must be continuous.

Here is the protocol:

FOR $n = 1, 2 \ldots$
Reality announces $x_n \in X$.
Skeptic announces continuous $S_n: [0, 1] \to R$.
Forecaster announces $p_n \in [0, 1]$.
Reality announces $y_n \in \{0, 1\}$.
Skeptic's profit $= S_n(p_n)(y_n - p_n)$.

The theorem I will now prove is that Forecaster has a strategy that keeps Skeptic from making any money in this protocol. Here is the strategy:

- If $S_n(p) > 0$ for all p, take $p_n = 1$.
- If $S_n(p) < 0$ for all p, take $p_n = 0$.
- Otherwise, choose p_n so that $S_n(p_n) = 0$.

Recall the intermediate value theorem: a continuous function is always positive, always negative, or has a zero. If S_n is always positive, then $p_n = 1$ makes Skeptic's

profit negative or zero no matter what Reality does. If S_n is always negative, then $p_n = 0$ makes Skeptic's profit negative or zero no matter what Reality does. If S_n has a zero, then a value for p_n such that $S_n(p_n)=0$ makes Skeptic's profit zero no matter what Reality does.

Universal tests and universal priors

A fundamental idea of game-theoretic probability is that a statistical test can be implemented by a betting strategy, a strategy for multiplying the capital you risk. Of course, we don't just want to pass one test; we want to pass a lot of tests. But as I will now explain, it is generally possible to merge the tests we want to pass into a single test.

If I have one strategy that risks just one dollar and turns it into an infinite sum whenever Forecaster's probabilities and Reality's outcomes fail to match in one particular way, and I have another strategy that will do the same thing whenever they fail to match in a different way, then what should I do? I should play 50 cents on the one strategy and 50 cents on the other, because multiplying 50 cents by infinity is just as good as multiplying a dollar by infinity: I get infinity either way. I get infinitely rich if either violation occurs.

Dividing your money between strategies is the same as averaging the strategies. You can use a weighted average if you want. You can average a countable number of strategies if you like. As long as you put a little bit of money on each strategy, you get infinity when that strategy multiplies its initial capital by infinity. And as Abraham Wald explained in the 1930s, when he was making sense of Richard von Mises's concept of a collective: no language allows us to devise more than a countable number of tests. In practice, we can devise only a finite number. So average them to get a portmanteau test – or *universal test* if we want to call it that.[17] Then Forecaster's problem comes down to beating a single test, and we just learned how to do that.

[17] Even in theory, we cannot really quite construct a universal test for a given language, because the countable set of tests we can devise in that language is not recursively enumerable.

As I explained earlier, a strategy for Forecaster is formally the same as a probability distribution P for the whole sequence $y_1, y_2, \ldots$ of moves by Reality; to get Forecaster's probability forecast for y_n in the nth round of the game, you condition P on $y_1, \ldots, y_{n-1}$.[18] So the strategy of defensive forecasting, when applied to a more or less universal test, will produce a more or less universal prior distribution for Reality's moves. As this suggests, what I have been explaining here is a finitary version of the notion of a universal prior distribution first explored by Leonid Levin in the 1970s.[19]

Calibration

The discussion has become too general, too abstract, too infinitary. Let me bring it back down to something specific, concrete, and finite. Suppose there is no extra information x_n, so that Forecaster's task is merely to predict y_n from $y_1, \ldots, y_{n-1}$ for $n=1,2,\ldots$. Suppose further that we are content to test Forecaster's probability predictions in a simple way: we want them to be calibrated.

Recall that a probability forecaster is *calibrated at* 30%, if you find that the event happened 30% of those times when you look at all the times when he said $p = 0.3$. Of course, to make this practical, we have to make it a little fuzzier: when you look at all the times when p is close to 30%, you want to see that the event happened about 30% of the time. We also want this to happen for 35%. We might ask for calibration for a couple of dozen different values of p, say $p = 0.05$, $p = 0.10$, $p = 0.15$, and so on, and also some values closer to 0 and to 1.

In Section 2, we saw Ville's law of large numbers for $p = 0.50$; this tells us how Skeptic can test Forecaster for calibration at 50%. The strategy $s_n(y_1, \ldots, y_{n-1})$ given there, used only on those rounds n where Forecaster puts p_n close to 0.50, will makes Skeptic rich if there are many such rounds and Reality does not set $y_n=1$ for about 50% of them. As I also mentioned in Section 2, Skeptic has a similar strategy

[18] To make this statement strictly correct in the case where $y_1, \ldots, y_{n-1}$ have zero probability, so that the conditional probability is not defined, we must go back to nineteenth-century practice, before measure theory became the official foundation for probability, and assume that P is given as a system of conditional probabilities, not as a single probability measure. See the final chapter of Ville (1939).

[19] Levin was the first to demonstrate the existence of distributions that withstand universal tests. Their existence is only theoretical, however, inasmuch as their values are not computable. The best mathematical treatment of the topic is Gács (2005). For a historical review, see Bienvenu et al. (2009).

for any other value of p. Let's merge the strategies for Skeptic for the two dozen or so values of p that we have chosen by averaging them. This gives us a strategy that makes Skeptic rich if he is uncalibrated for any of the values of p (or uncalibrated anywhere, because we count any p as close to one of these values).

Vovk developed this idea a little further, using many more values of p and looking at what happens to the average strategy as the number goes to infinity. He obtained a very simple expression for the resulting strategy for Skeptic at each round n of the game:

$$S_n(p) = \sum_{i=1}^{n-1} e^{-C(p-p_i)^2} (y_i - p_i).$$

The constant C is a learning rate, which you can choose more or less as you please.

This strategy uses all the information in the game so far. Remember that the y_i are Reality's outcomes, 0s and 1s, while the p_i are Forecaster's predictions. If you average the discrepancies between prediction and outcome, $(y_i - p_i)$, you want the average to be close to zero; this was our martingale law of large numbers. But here we have a weighted average, using the Gaussian kernel

$$e^{-C(p-p_i)^2}.$$

This kernel gives a weight of almost zero to any round i for which p_i is not close to p, so that you're averaging only over rounds where p_i is close to p. Any other kernel with this feature would work about as well.

Now remember that Forecaster uses the function $S_n(p)$, which defines a strategy for Skeptic, to obtain his own strategy. He defends against S_n by choosing p so that $S_n(p)=0$. In other words, he uses a value of p where he did best in the past. This is what most of us do in life. What game are you going to play? You play the game you're good at!

In order to pass statistical tests in the on-line setting, you have to keep any trend from emerging, and the way to keep any trend from emerging is to repeat what has been working for you. When you do this, the worst that can happen is that your

good performance in that area will be a little less good, probably not bad enough to fail a test. But if you keep venturing into territory where you have done badly, Reality has a chance to make your bad performance so far into a really bad trend, making you fail a test.

As I mentioned earlier, there are virtues a Forecaster might aspire to in addition to calibration. You might also want the convergence to happen the right way, the way prescribed by the law of the iterated logarithm. We've got a strategy for the law of the iterated logarithm too; you can average that in if it's important to you, though you might need a lot of trials for it to make any difference.

More importantly, in applications we need to bring the x_n back, to make sure that Forecaster has good *resolution* as well as good calibration. This means that you look not just at the times in the past when Forecaster said the probability of rain was 30%, but you also look more specifically at those times when he said 30% just after it had rained the day before. It should rain 30% of these times. And so on for other important side information $x_1,\ldots,x_n$. You can get resolution, within reason, in the same way that we get calibration; you need a kernel that combines the comparison of p_i to p with a comparison of x_i to x_n.

We already knew that probability could be estimated from a random sample. I've told you something less well known: In the on-line setting, we don't need the sample to be random in any sense in which samples can be unrandom. We can pass statistical tests regardless of how Reality behaves.[20]

On-line, everything is stochastic

Having been trained as a mathematical statistician in the 1970s, I have long been familiar with the concept of a stochastic process with unknown probabilities that we have to estimate as we go along. But the success of defensive forecasting has taught me something about this concept that I did not understand before. I had thought that we needed to know something about reality, or at least correctly guess something about reality, in order for this *estimation as we go along* to succeed. I thought that we needed a correct model of reality. Defensive forecasting has taught

[20] That's why I call this player Reality, not Nature. Nature follows laws. Reality can play as he pleases.

me that this is less true than I had thought.[21] We don't need a correct model of reality. We can succeed no matter how reality behaves.

I had thought that sometimes a process might not be stochastic. Sometimes there might be no way to give probabilities that would pass statistical tests. Now that I have learned that I was wrong, I want to say that everything is stochastic, but I also want to say that this statement has no empirical content. It is not falsifiable in the on-line setting.

5. Broader perspectives

This presentation has been about the on-line setting. There, I have argued, good probability forecasting is possible no matter how reality behaves. But we are not always on-line. I want to conclude with a word about other domains of experience where we can use mathematical probability. I see two very large domains: causality and judgement.

Probabilistic causality

Sometimes we do know something about reality's behaviour. Sometimes we have rules for making probability forecasts that will pass statistical tests but do not depend on the particular sequence of events we have chosen to consider. For an empiricist, it is reasonable to call our knowledge of these rules for prediction causal knowledge.

Probabilistic causality is a vast topic. I cannot even begin to explore it here, but I want to mention that I once tried (see Shafer, 1996). There I used the same game-theoretic notion of probability I have used in this talk: probability forecasts are offers to bet that are tested by strategies for multiplying the capital risked. These offers may fall short of complete probability forecasts, so that we obtain only upper and lower probabilities, as in my 2001 book with Vovk and in the framework that Walley (1991) dubbed, misleadingly in my view "imprecise".

[21] Well, OK, in practice there is still some truth to it. We need to get the kernel right – that is, to identify the features of our information that are important for prediction.

On the philosophical side, the challenge here is to reconcile realists with the thesis that causal talk is primarily talk about regularities in successful prediction. No one ever has the last word in such philosophical arguments, but once the game-theoretic understanding of mathematical probability is accepted, it becomes clear that probability can only be about predicting phenomena, not about "generating" them.

Probability judgement

The on-line setting does not start with probabilities, but it starts with a structure that turns out to be almost as powerful: a sequence of experiments. I have told you how to make up a probability for a particular event, but only when this event was embedded in a sequence. Everyone had agreed, I assumed, on this sequence.

In real life, people often don't agree on the sequence in which to place an event. I have a dispute with my neighbour, we go to trial, he puts what happened in his sequence of what has happened to him, and I put what happened in my sequence of what has happened to me. We go to the judge, each of us argues for the relevance of his sequence. The judge's job is to decide what is relevant. Are the bets justified by one of the sequences still justified when we see the information in the other sequence? This is a matter of judgement, and it can sometimes be framed as a judgement that betting offers that cannot be beat in one sequence still cannot be beat using the additional information in the other sequence.

This line of thought leads to Bayesian conditioning and Dempster–Shafer theory (see Shafer, in press). Again, you may decide that only some betting offers remain reliable, and therefore settle for upper and lower probabilities. But this is surely a topic for another day.

Acknowledgements

I am grateful to the Cournot Centre for the opportunity to present these ideas in this format, the informality of which may lower some of the barriers to understanding built into more established modes of publication. In revising the transcript of my talk, I have benefitted from questions and comments from a number of colleagues, especially Michel Armatte, Darrell Rowbottom, and Volodya Vovk.

References

Cesa-Bianchi, Nicolò, and Gábor Lugosi (2006), *Prediction, Learning, and Games*, Cambridge: Cambridge University Press.

Bienvenu, Laurent, Glenn Shafer and Alexander Shen (2009), "On the history of martingales in the study of randomness", *Electronic Journal for History of Probability and Statistics*, 5(1) (www.jehps.net).

Dawid, A. Philip (1984), "Statistical theory: the prequential approach", *Journal of the Royal Statistical Society A*, 147, pp. 278–92.

Dawid, A. Philip, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin and Vladimir Vovk (2011), "Insuring against loss of evidence in game-theoretic probability", *Statistics and Probability Letters*, 81 pp. 157–162.

Drouet, Isabelle (in press), "Propensities and conditional probabilities", *International Journal for Approximate Reasoning*.

Gács, Péter (2005), "Uniform test of algorithmic randomness over a general space", *Theoretical Computer Science,* 341, pp. 91–137. This article is strengthened and extended in Gács' *Lecture notes on descriptional complexity and randomness*, available from his personal web page.

Hacohen, Malachi Haim (2000), *Karl Popper: The Formative Years, 1902–1945*. Cambridge: Cambridge University Press.

Hennig, Christian (2007), "Falsification of propensity models by statistical tests and the goodness-of-fit paradox", *Philosophia Mathematica*, 15(2), pp. 166–192.

Kolmogorov, Andrei N. (1933), *Grundbegriffe der Wahrscheinlichkeits-rechnung*, Vienna, Austria: Springer. English translation, *Foundations of the Theory of Probability* (1950, 1956), New York, NY: Chelsea.

Kolmogorov, Andrei N. (1951), "Probability (in Russian)", *Great Soviet Encyclopedia*, 7, Second Edition, pp. 508–10.

Neyman, Jerzy (1960), "Indeterminism in science and new demands on statisticians", *Journal of the American Statistical Association*, 55, pp. 625–39.

Popper, Karl R. (1935), *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*, Vienna, Austria: Springer. English edition, *The Logic of Scientific Discovery*, (1959), London: Hutchinson.

Popper, Karl R. (1938), "A set of independent axioms for probability", *Mind*, 47, pp. 275–77.

Popper, Karl R. (1982), *The Open Universe: An Argument for Indeterminism*, Volume 2 of *Postscript to the Logic of Scientific Discovery*, Totowa, NJ: Rowman and Littlefield.

Popper, Karl R. (1983), *Realism and the Aim of Science*, Volume 1 of *Postscript to the Logic of Scientific Discovery*, Totowa, NJ: Rowman and Littlefield.

Psillos, Stathis (1999), *Scientific Realism: How Science Tracks Truth*, London: Routledge.

Shafer, Glenn (1996), *The Art of Causal Conjecture*, Cambridge, MA:The MIT Press.

Shafer, Glenn (2007), "From Cournot's principle to market efficiency," in Jean-Philippe Touffut (ed.), *Augustin Cournot: Modelling Economics*, Cheltenham, UK, and Northampton, MA, USA: Edward Elgar, pp. 55–95.

Shafer, Glenn (in press), "A betting interpretation for probabilities and Dempster–Shafer degrees of belief", *International Journal for Approximate Reasoning*.

Shafer, Glenn and Vladimir Vovk (2001), *Probability and Finance, It's Only a Game*, New York, NY: Wiley. For sample chapters and related working papers, see www.probabilityandfinance.com.

Shafer, Glenn, Alexander Shen, Nikolai Vereshchagin and Vladimir Vovk (in press), "Test martingales, Bayes factors, and p-values", *Statistical Science*.

van Fraassen, Bas C. (1980), *The Scientific Image*, Oxford: Oxford University Press.

Ville, Jean André (1939), *Etude critique de la notion de collectif*, Paris: Gauthiers-Villars.

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