Center for
Advanced Study in
the Behavioral Sciences

Annual Report
Year Ending August 31,
1982

ESSAY

How Is the Past Related to the Future?

Joel E. Cohen

The following essay is based on a talk given on April 9, 1982, to the Center's Board of Trustees.

A story is told of two men talking. One says, "Tell me, why do you always answer a question with a question?" The other says, "Yes, why?"

The short answer to the question "How is the past related to the future?" is "I don't know." A longer answer, which follows, replaces this simple question with not one but two other questions. These more complicated questions show why we cannot be certain how the past is related to the future.

Before asking these two questions, let me provide a framework for them by making two assumptions. The major assumption is that the universe evolves through mechanisms that act locally in time. On this assumption I shall consider as possible explanations of how the past is connected to the future only mechanisms or models or hypotheses that link the state of the world at one instant to the state of the world at the next instant.

This assumption is intelligible (whether or not it is true) if time is taken as discrete or quantized, since in that event after any instant there is a unique next instant. If, by contrast, we take time as continuous, like the position of a particle moving on a line, a tedious technical detour, which I propose to avoid here, would be necessary to define an infinitesimal increment of time in some sensible way. So my second, but minor, assumption is that time is discrete or approximately so.

To summarize, I assume that the vast sweep of cosmic history must be generated by the repeated iteration of some process that moves the universe from one instant to the next.

By what reasoning do I support this assumption? Whatever influence the past has on the future passes through the bottleneck of the present. So the state of the universe now at any given instant embodies all the influence (whether deterministic or probabilistic) that the past will have on the state of the universe an instant later. When that future instant becomes the present, it too embodies the influence of all prior history on the next following instant. Unless you imagine that causality leapfrogs, i.e., that the past somehow influences the future by means that are not detectable in the present state of the universe, you should share my assumption that the trajectory of the world is the large-scale result of concatenated processes acting from one time-point to the next. In the same way, wind and water, eroding one speck of stone after another, carved the Grand Canyon.

Within this framework, we can hope to understand how the past is related to the future only if we can confidently answer "Yes" to both the following questions:

(1) If we observe the world for a long period, can we infer the elementary process by which the world changes from one instant to the next? More technically, given a sample path, can we infer its infinitesimal generator? Given the Grand Canyon, can we guess the physical laws that govern the sculpting of grain after grain, speck after speck?

(2) If we suspect or are told what the elementary process is by which the world changes from one instant to the next, can we describe the world's long-run behavior? Given the physical principles of erosion, could we foresee the Grand Canyon?

I see no reason to believe that the answer to either question is "Yes." In the rest of this essay, I will try to persuade you, by means of three examples. These three examples are toy models of the world. They are all familiar to mathematicians who specialize in the areas from which the examples come; like lions and tigers in the care of zookeepers, these examples have lost their terrors for the specialists. But when these mathematical beasts escape their usual confines, their power and cunning return. With an understanding of these three examples will come, I hope, a clearer sense of the difficulties involved in understanding how the past is related to the future.

Example 1

Consider the plight of Sam (Figure 1). He wants to buy a hamburger. He sees that in the past 400 zillion of Big M's have been sold, while approximately 17 of Joe's have been sold. Should he infer that the difference in sales reveals some real advantage of Big M over Joe (superiority of product, location, or advertising), or could the difference in sales be due to chance alone?

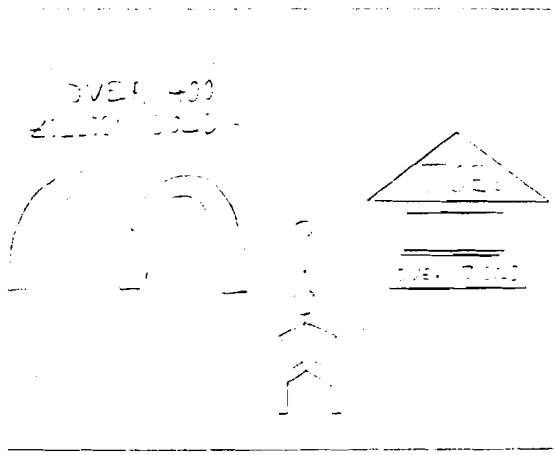

FIGURE 1
For a long time, Big M has had the lead over Joe in the number of hamburgers sold. Could this sustained lead be due to chance alone?

Let us formalize the problem, forgetting about hamburgers. Suppose we have two players A (Big M) and B (Joe). Let the times $t = 1$, $t = 2$, $t = 3$, etc., be the times when the first, second, third, etc., customer chooses to buy from Big M or from Joe. We will call each such choice a trial, and A wins the trial and B loses if the customer chooses Big M. For each time t, keep track of the score by letting $X_t = 1$ if A wins and B loses and

$X_t = -1$ if A loses and B wins. We define A's cumulative score at time t to be $S_t = X_1 + X_2 + \ldots + X_t$, that is, the sum of the scores at all times up to and including t. Thus A has won more trials than B has by time t if S_t exceeds 0; A and B are tied at time t if $S_t = 0$; and A has won fewer trials than B by time t if S_t is less than 0. One can graph the progress of competition between A and B as shown in Figure 2, with time on the horizontal axis and the cumulative score on the vertical axis.

For convenience, I propose to ignore ties and to say "A leads" provided that the broken line segment connecting successive values of S_t lies above the horizontal axis, while "B leads" provided the graph of S_t lies below the horizontal axis. Thus, in Figure 2, A leads six times in eight trials, because the line is above the t axis for six of the eight time intervals shown.

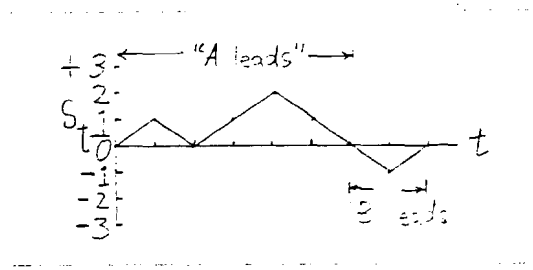

FIGURE 2
An illustrative graph of the cumulative score of player A (vertical axis) as a function of the number t of trials (horizontal axis). When the graph is above the t axis, A leads by definition; when below, B leads.

Now suppose that at each trial, player A wins with some fixed unknown probability, which I call p. Here p is a number greater than 0 and less than 1. Player B wins with probability $1 - p$. Suppose also that all trials are mutually independent, so that regardless of how many times A has won or lost in the past, A's probability of winning at the next trial is still just p. Think of A's winning or losing as being determined by successive tosses of a coin with enough wear on one side to make the probability of heads (say) just p.

Note that this mathematical model for the sequence of choices between Big M and Joe is highly idealized. For example, real people might try to influence their friends to buy the kind of hamburgers they like; or preferences might change over time. The model excludes all such complexities. I want to show that our intuition may have difficulty even with a simple process.

In the setting of this hypothetical sequence of trials (assuming independence and a constant p), I ask: How good are we at making inferences about p from current information about the fraction of time that A leads B (i.e., from current information about the fraction of time that the graph of S_t lies above the t axis)?

More specifically, suppose that after $t = 20$ trials, one of the two players (it doesn't matter which one) has been observed *always* to be in the lead. Mathematically we suppose either that $S_1 > 0, S_2 \geq 0, S_3 > 0, \ldots, S_{20} \geq 0$, or that $S_1 < 0, S_2 \leq 0, S_3 < 0, \ldots, S_{20} \leq 0$. This is *not* supposing that one or the other player won at every trial (*not*: $X_t > 0$, $t = 1, 2, \ldots, 20$, or $X_t < 0$, $t = 1, 2, \ldots, 20$), but just that whoever took the lead on the first trial kept the lead through all 20 trials.

Is this evidence against $p = \frac{1}{2}$? In other words, among sequences of 20 trials with $p = \frac{1}{2}$, is it rare for whoever leads after one trial never to lose the lead to the other player? If people were choosing between Big M and Joe independently of each other, by flipping a fair coin with probability $\frac{1}{2}$, would we be surprised to see that whoever took the lead after the first customer made her choice always kept the lead?

To gain time to think about your answer, and to help develop your intuition, consider sequences of only two trials (Figure 3). All four possible paths of the graph of S_t are shown. It is impossible for any of them to cross the t axis after only two trials. Thus after two trials it is not rare, but certain, that whichever player takes the lead after the first trial keeps the lead on the next trial (by our definition of "A leads").

Now what about 20 trials? In this case, the graph of S_t can cross the t axis. Among all possible graphs of S_t over 20 trials, would it be rare to find a graph that never crossed the t axis? More quantitatively, in what fraction of sequences of 20 trials with $p = \frac{1}{2}$ would you expect one player or the other always to lead?

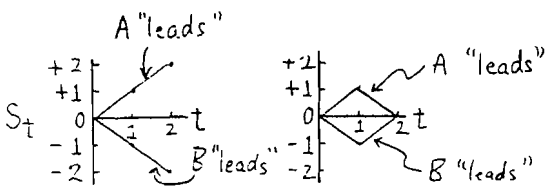

FIGURE 3
In sequences of only two trials, it is not possible for the lead to change from one player to the other.

Please stop here, and make a guess before you read the answer in the next paragraph.

Whichever player gets the lead on the first trial keeps that lead in more than 35 percent of sequences of 20 independent trials with $p = ½$, so more than one in three graphs of S_t would never cross the t axis.

After 20 trials, let k be the number of times the luckier player leads. In case the two players have each won the same number of times, let us put $k = 10$. Otherwise, k is just the number of line segments in the graph of S_t that lie on the side of the winner. E.g., in Figure 2, A leads six times.

What is the most probable value of k, assuming that $p = ½$? To put the question in a more leading form, if $p = ½$, is it most likely, among all possible sequences of 20 trials, that each player would lead for ten trials?

What is the least probable value of k?

Is it more likely that $k = 12$ or that $k = 18$? If $p = ½$, is it more likely that the luckier player will lead by not very much ($k = 12$) or by a lot ($k = 18$)?

The most probable value of k is 20. The least probable value of k is 10. The value $k = 18$ is more probable than $k = 12$.

I hope you are at least slightly surprised by these answers. But you might feel they arise at least in part because of the small number of trials. After all, when there are only two trials, the geometry alone obliges one player to lead twice, regardless of p.

To develop our intuition and avoid wear and tear on our coin-flipping fingers, I have asked a computer to simulate 100 independent trials with $p = ½$ and to draw a graph of S_t (on the vertical axis) as a function of t (on the horizontal axis). Figure 4 shows six such simulations. The horizontal axis is moved up and down in the different panels of Figure 4 in order to center the graph of S_t. The main feature of these pictures is that the graph usually falls quite lopsidedly above or below the t axis. In other words, one player usually keeps the lead most of the time; the lead does not appear to be evenly distributed between the two players.

Still, 100 trials is a relatively small number. With 1,000 trials, the distribution of the lead might even out. Figure 5 shows three such simulations. The tick marks that show individual trials on the t axis are so close together they are not distinguishable. Once again, the lead seems to pass to one player or the other and to remain there in a most lopsided way. (These simulations appear exactly as they came off the computer. I did not select them for "typicality" or any other feature.)

For truly large numbers of trials, the computer is a less practical instrument of understanding than theory. Here is what theory tells us when the number of trials (customers choosing hamburgers) gets very large, e.g., larger than ten thousand.

Suppose that successive trials are independent, that the probability of player A's winning any given trial is an unknown constant p, and that at the end of 10,000 or more trials, one of the players is observed to have had the lead 99.4 percent of the time. Does this suggest that p is not ½?

No. If $p = ½$, one player would keep the lead 99.4 percent of the time in roughly one of ten sequences of trials, as long as the number of trials were large.

If trials occurred at a rate of one per second for 365 days, and if $p = ½$, how long would the lead time of the more fortunate player have to be for the event to have probability of one in one hundred (a level of improbability commonly used by behavioral scientists to establish "statistical significance")? In other words, name the

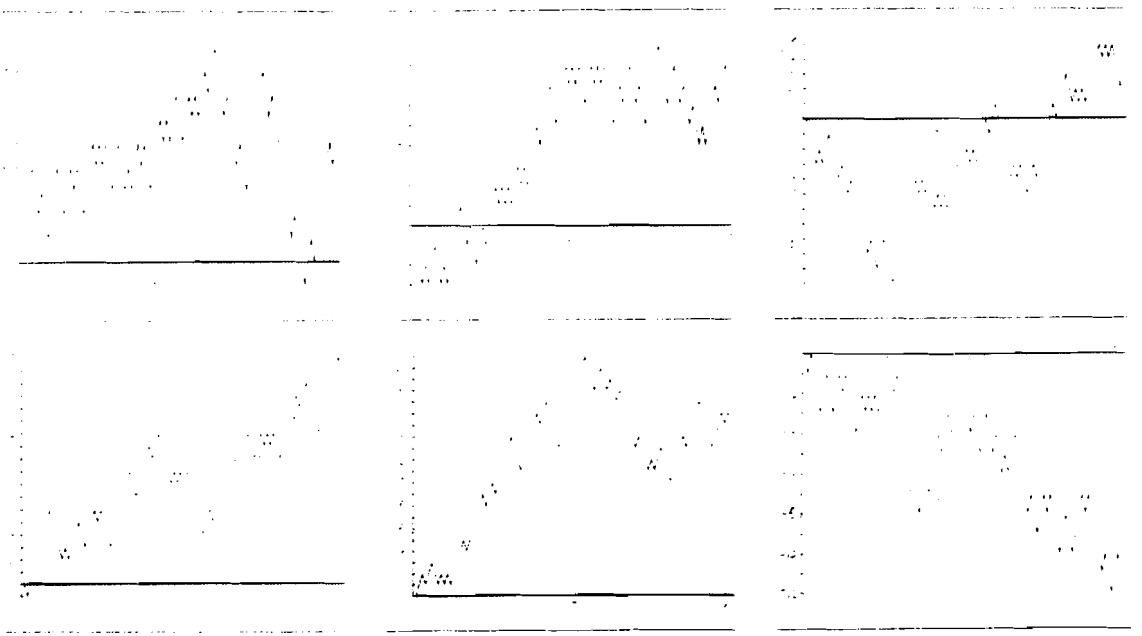

FIGURE 4
Six independent simulations, each of 100 independent trials, with A's probability of success $p = \frac{1}{2}$. The axes are as in Figure 2.

time T such that if the lead time of the more fortunate player were to equal or exceed T, you would be willing to reject the null hypothesis of $p = \frac{1}{2}$ at the 0.01 level.

You should take $T = 364$ days, 23 hours, 27.6 minutes, approximately, or the entire year except 32.4 minutes. If the more fortunate player leads for any period shorter than this T, e.g., merely 364 days and 12 hours, it is not evidence, statistically significant at the one percent level, against $p = \frac{1}{2}$. The advantage of the more fortunate player could be due to chance alone.

These numerical examples illustrate what is known as the arc sine law. The arc sine law was discovered by Paul Lévy in 1939 and was greatly extended by Paul Erdös and Mark Kac in 1947. It represents an understanding of the nature of random fluctuations that was simply unknown to earlier centuries. Formally, the arc sine law states that the probability density function of the fraction x of times that player A leads when $p = \frac{1}{2}$ in a long series of independent trials is

$$f(x) = 1/\{\pi[x(1-x)]^{1/2}\}.$$

The qualitative shape of the graph of $f(x)$ as a function of x is shown in Figure 6. The point is that $f(x)$ becomes very big as x approaches 0 or 1; it is most likely that A will lead either all of the time or none of the time. The low point of $f(x)$ occurs at $x = \frac{1}{2}$, meaning that it is least likely that A shares the lead evenly with B.

The arc sine law shows that the fraction of time that one player leads the other must be extremely close to 1 before it provides evidence that, on the next trial, one player has a better chance than another.

The probabilist William Feller, to whose great book I owe these numerical examples (without some of the frosting) as well as my understanding of the arc sine law, comments on these results: "If even the simple coin-tossing game [our model for the customers' choices] leads to paradoxical results that contradict our intuition, the latter cannot serve as a reliable guide in more complicated situations."

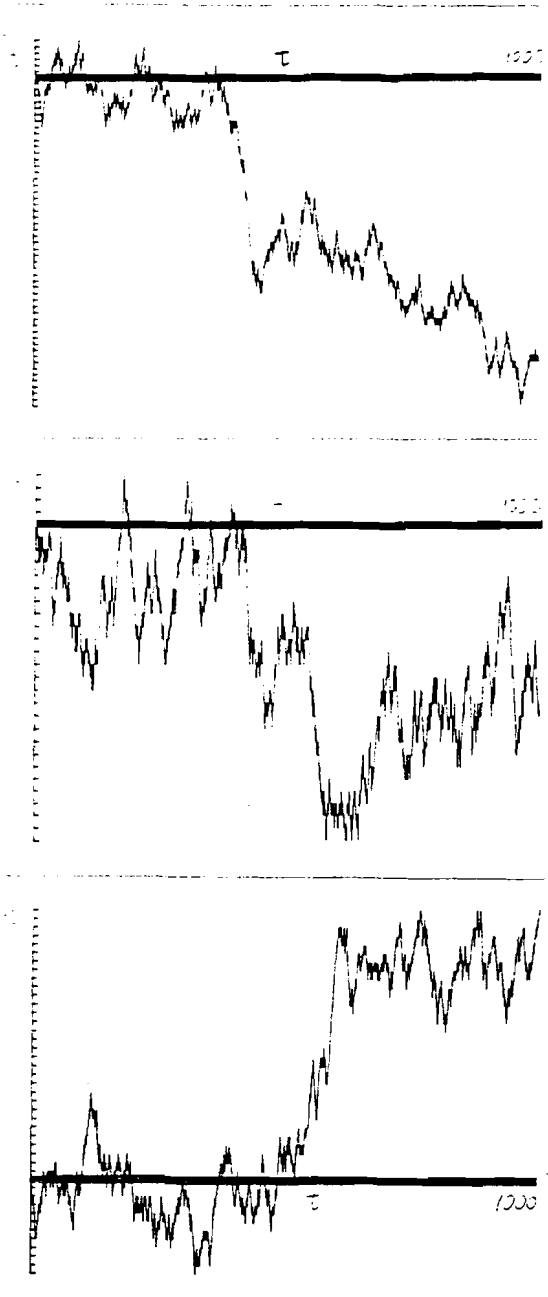

FIGURE 5
Three independent simulations, each of 1,000 independent trials, with A's probability of success $p = ½$. The axes are as in Figure 2.

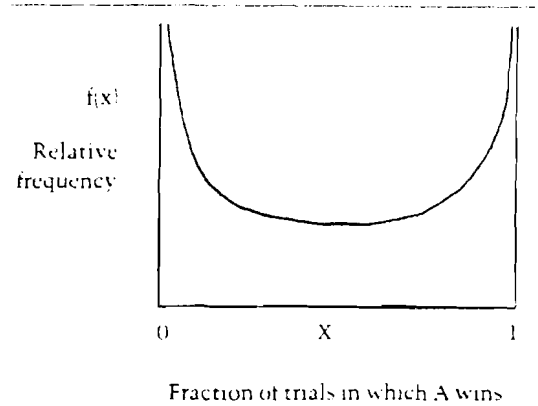

FIGURE 6
A qualitative graph of the probability density $f(x)$, or relative frequency, with which player A leads for a fraction x of trials, in very long sequences of trials ($0 < x < 1$), assuming independent trials and a probability $p = ½$ of A's winning on each trial. It is least likely that A will lead exactly ½ the time and most likely that A will lead either never or all the time.

Now, with the same model (independent and identical trials, with probability p that A wins on each trial), consider another summary of the past: the ratio of the number of trials so far on which A has won to the total number of trials so far. Mathematicians have long known that, except for sequences of trials so rare they may safely be ignored, the ratio just defined rapidly gets closer and closer to p as the total number of trials gets larger and larger. So if p is different from ½, summarizing the past by the fraction of trials on which A has won (or Big M gets the customer) will rapidly reveal that p is not ½. Rigorous statistical theory prescribes how to test whether the observed fraction of trials on which A wins rejects the possibility that $p = ½$.

With this ratio as a summary of the past, we then have nearly complete information about future trials: After a large number of trials, the probability that A will win on the next trial is very close to the fraction of times A has won so far. Using this summary leads to none of the sur-

prises that accompany use of the fraction of past trials on which A has been in the lead.

We draw two conclusions.

(1) If you pick the right summary of prior experience, you derive the right information from the past about the future.

(2) To pick the right summary, it helps if you already know how the past and future work.

To reinforce conclusion (2), we give next an example in which the ratio summary that works so nicely above behaves bizarrely.

Example 2

Imagine a dish for growing bacteria that starts out with just two bacteria. Suppose that the bacteria are identical in all respects save that one is round and one is square (or one is black and one is blue, or one has hair and one is bald, etc.).

As is well known, in suitable environments bacteria like to divide. Suppose that when these bacteria divide into two daughter cells, the daughter cells are equally and fully ready to divide again, but that the interval between the birth of a cell and its division into two daughters is random, independently and identically distributed for all cells. The cells never die.

Let $t = 1$ be the time the first division occurs in the dish; $t = 2$ the time the second division occurs; and so on. Since the two original bacteria are assumed identical in their propensity to divide, the round one is as likely to divide first as the square one. With probability ½ the dish will contain two round bacteria and one square one just after $t = 1$, when the first division is completed, and with probability ½ the dish will contain two square bacteria and one round one just after $t = 1$. The possibilities are illustrated in Figure 7.

Let us consider the possibilities that follow from the first case through the next division that occurs. If there are two round bacteria and one square one just after $t = 1$ (the left half of Figure 7), then there are two chances in three that the next division will occur in a round cell, and only one chance in three that the square cell will be next to divide.

What will happen to the fraction of round bacteria in the dish after many divisions have occurred? Please guess.

Here are some of the answers people have given me when I have asked them to guess:

(a) If the first cell to divide is a round one, then ultimately the fraction of round bacteria approaches 1 and the fraction of square bacteria becomes vanishingly small; if the first cell to divide is square, the reverse happens.

(b) Regardless of what happens on the early trials, the fraction of round bacteria in the long run settles down near or at ½.

(c) Since there will always be both round and square bacteria in the box, even when the round cells are in the overwhelming majority, the square cells can always come back, so the fraction of round cells never settles down to any definite number but drifts aimlessly (nay, forlornly) between 0 and 1.

Perhaps you have some proposal not approximated by these three guesses. More likely, you

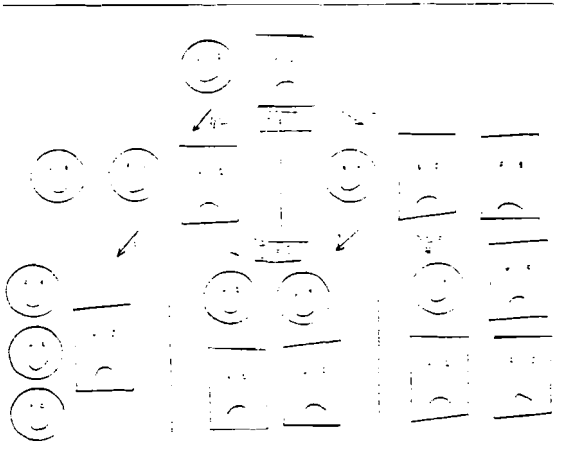

FIGURE 7
A dish starts out (at time $t = 0$) with two bacteria, one round, one square. Whenever a division occurs, the cell that divides is equally likely to be any of the cells already in the dish. Shown are the possiblities and conditional probabilities of each transition, for the first ($t = 1$) and second ($t = 2$) divisions.

wonder what these bacteria have to do with the behavioral sciences.

So consider instead the early days of the typewriter industry before the layout of letters on the keyboard became fixed. Imagine that two typewriter manufacturers introduced typewriters that were identical in all respects except the arrangement of letters. Call the two brands of typewriters A and B.

Suppose that two typewriters, one of each type, were sold or given away, and that people subsequently decided which type to buy by randomly visiting someone who had a typewriter, hearing the glories of the keyboard arrangement, and buying the type owned by the person they visited. (It is my impression that many decisions about which personal computer to buy are made in just this way.)

If this elementary process, which connects the fraction of type A keyboards after t purchases to the fraction of type A keyboards after $t + 1$ purchases, is repeated over and over again, what will happen to the fraction of type A keyboards sold in the long run?

In the case of keyboards for the Roman alphabet, we know that a single arrangement of letters, with minor variations for different languages, has become nearly universal (I think). Is this the only outcome consistent with the elementary imitative process I have hypothesized?

To convince you that other outcomes are at least possible, consider light switches. In the United States and many other countries, down is off. In Australia and New Zealand, down is on! *If* new countries chose the orientation of their light switches by choosing randomly among the practices of all existing countries (how they *actually* do it, I don't know), we would have an example where "down is off" had become widespread, but not universal.

Bacteria, keyboards, and light switches are

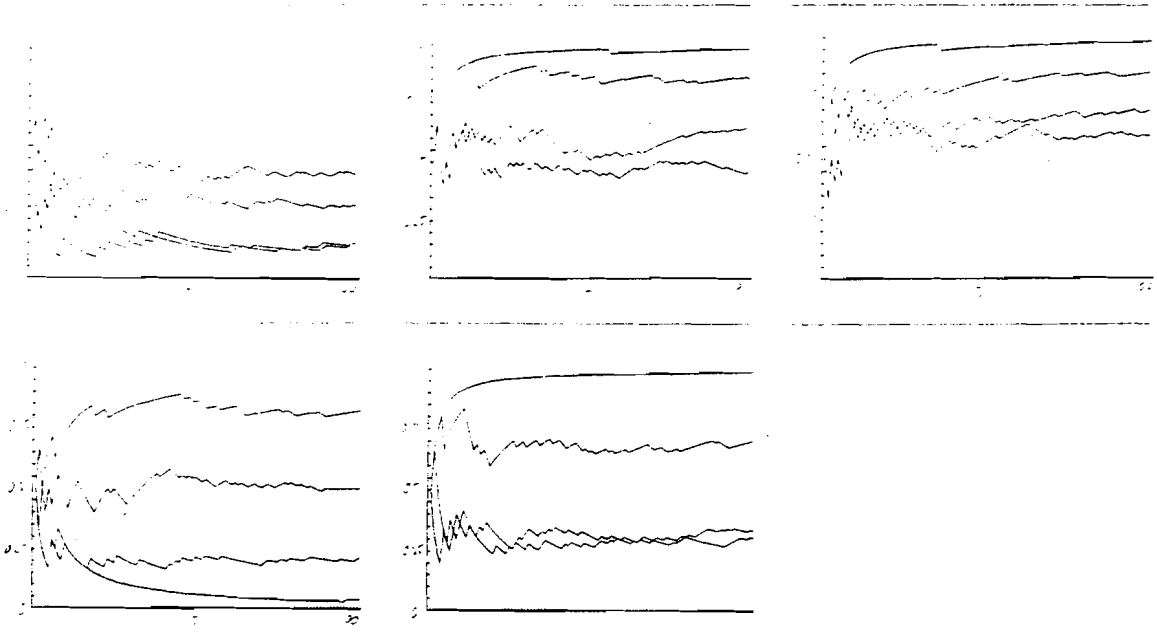

FIGURE 5
Each panel shows four independent simulations of the fraction of A balls (vertical axis) in an urn as a function of the number of balls added (horizontal axis), for the first 100 additions. On each trial, a ball is randomly picked out and then returned to the urn along with an additional ball bearing the same letter, A or B.

all interpretations of a so-called urn model. The urn model was described by F. Eggenberger and George Pólya in 1923. A careful analysis was published by David Blackwell and David Kendall in 1964.

We have a large urn with one ball marked A and one ball marked B. At each time $t = 1$, $t = 2$, $t = 3$, etc., we stir the balls in the urn, choose one at random, note the letter on it, replace the ball, and add another ball with the same letter on it.

After a large number of drawings, replacements, and additions, what happens to the fraction of A balls?

I did not guess the correct answer when this question was put to me, and no one to whom I have put it, including mathematicians and specialists in probability and statistics, has guessed correctly.

To sharpen our intuition, we again have recourse to the computer to simulate the random drawing of a ball marked A or B from an urn and its replacement in the urn along with an additional ball bearing the same letter. In each panel of Figure 8, the horizontal axis indicates the number of balls that have been added to the initial two balls in the urn (starting from 0 additional balls), and the vertical axis shows the fraction of balls in the urn marked A (starting from ½, when no balls have been added). In each panel, four independent simulations, each of 100 additional balls, are shown. In all the simulations it appears that the fraction of A balls fluctuates at first, then levels out; but *where* the fraction of A balls levels out for one simulation appears to have no connection to *where* the fraction of A balls levels out for any other simulation.

To see whether the fraction of A balls will eventually flatten out, will wander, or will converge on ½ or some other number, I have simulated the addition of 1,000 balls to the urn. Each panel of Figure 9 shows four independent simulations of 1,000 draws and additions. In all cases, the fraction of A balls fluctuates less and less as the number of draws increases, but the fraction of A balls after 1,000 balls have been added seems to vary wildly from one simulation to another.

What is happening here?

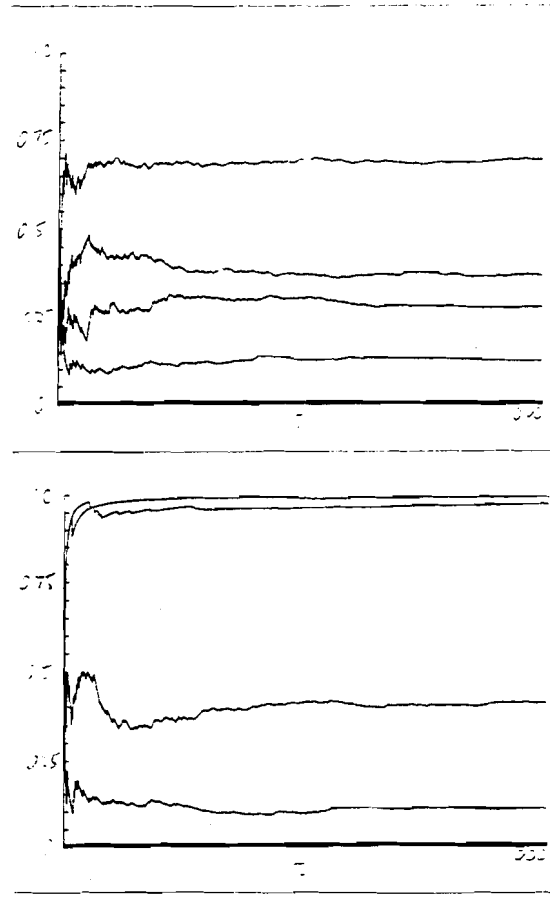

FIGURE 9
Each panel shows four independent simulations of the fraction of A balls as 1,000 balls are added to the urn. Axes are as in Figure 8.

Theory shows that if I follow the rules of drawing at random, returning the ball along with one of the same kind drawn, then the fraction of A balls in my urn will gradually settle down to a fixed number between 0 and 1, possibly including 0 or 1, with smaller and smaller fluctuations as time goes on. The fraction of A balls approaches a limit in the sense of elementary calculus. However, if you follow the same rules, the fraction of A balls in your urn will also settle down to some fixed number between 0 and 1, but the fraction for your urn is totally independent of the ulti-

mate fraction of A balls in my urn! In fact, if many replicas of the urn model are run simultaneously for a long time, theory shows that the ultimate fraction of A balls in any given urn is just as likely to lie between 0 and 0.1 as it is to lie between 0.45 and 0.55 or between 0.9 and 1.0. The ultimate fraction of A balls is uniformly distributed between 0 and 1.

This long-run behavior of the fraction of A balls is an example of *regularity without reproducibility*. In statistical jargon, we call it an example of almost sure convergence to a nondegenerate limit random variable.

To jump from urn to universe, suppose the universe we live in is not unique, but is one of many replicated universes started from the same initial conditions and governed by a chance mechanism analogous to that of the urn. In this universe, some measured variable corresponding to the fraction of A balls, say, the fraction of iron among all the elements, at first fluctuates but then settles down to some long-run level. We interpret this long-run regularity as a law of nature and construct a physical theory to explain it. Meanwhile, in the universe next door, unknown to us, the fraction of iron among the elements is settling down to some completely different level and the creatures who live next door (if there are any—perhaps the presence of life is another random fluctuation) are trying to explain that level as a law of nature.

If, in fact, there were some law governing the probability distribution of the fraction of iron among the elements over the entire ensemble of replicated universes, it would be quite difficult to guess that law by knowing the single simulation revealed through the universe you and I are confined to.

To make life more difficult, suppose the asymptotic limit of the fraction of A balls in our particular urn is going to be p, some number between 0 and 1. Then theory tells us that our sequence of draws will look as if, instead of drawing balls, I were repeatedly adding an A ball with probability p and a B ball with probability $1 - p$, without paying any attention whatever to what is in the urn at the moment. No conceivable statistical test could distinguish between the actual process used in the urn model and independent choices of A and B balls with probabilities p and $1 - p$.

In the urn model, each random choice early in a sequence of draws shapes the average or expected long-run proportion of A balls in the indefinite future. Before any balls are drawn, when the fraction of A balls is ½, the average long-run fraction of A balls is also ½ (which is just the average of a uniform distribution between 0 and 1). If, after the first ball is added, the fraction of A balls is ⅔, then the average long-run fraction of A balls for all sequences of draws that begin this way is just ⅔; and so on.

If we think of each draw of an A ball as success on a trial, then we can compare the long-run behavior of the fraction of successes predicted by the coin-tossing model in Example 1 with that resulting from the urn model here. In coin tossing, the fraction of successes rapidly approaches a constant (usually ½ for a fair coin), and this constant is the same in every replication or simulation of the process. In the urn model, the fraction of successes again approaches a constant, but that constant fluctuates randomly from one replication to another and is strongly affected by the random events that occur early in the process.

Comparison of Examples 1 and 2 shows there is no reason to believe the notion that "things will average out in the long run." Whether things will average out in the long run depends on the details of how chance works, not merely on the presence of chance.

Example 3

We now come to a process that involves no chance. This example illustrates that surprises in the relation between past and future do not depend on probabilistic mechanisms. The process is known to demographers as the "components method" of population projection, and the property of this process that is obvious only to the initiated is known as the strong ergodic theorem of demography. Let me attempt a short account of these mysteries.

Consider a human population in a defined region. On January 1 of year t, e.g., $t = 1982$, let

$y_1^{(t)}$ be the number of females who will be one year old on their next birthday; let $y_2^{(t)}$ be the number of females who will be two years old on their next birthday; and generally let $y_i^{(t)}$ be the number of females who will be i years old on their next birthday, for $i = 1, 2, 3, \ldots, 120$. For brevity, I say that $y_i^{(t)}$ is the number of females in age class i in year t. The counting stops when there is no one left at that age to count, and the choice of 120 as the oldest age is arbitrary. I call the whole set of these 120 numbers $\{y_i^{(t)}\}$ a census.

How is this year's census $\{y_i^{(t)}\}$ related to last year's census $\{y_i^{(t-1)}\}$?

It will simplify the story, while not grossly affecting its outcome, if I assume at this point that the population experiences no emigration or immigration. In the real world, of course, migration can enormously alter a population's history. Here I exclude it to make a theoretical point.

In the absence of migration, the population can change only through births and deaths. One way to measure births and deaths is in terms of age-specific rates. For example, the birth rate of 25-year-old women is the number of children born in one year (or some other specific interval) per thousand women of that age. (Exactly when during the year the 25-year-olds are to be counted is a technical problem best left for the entertainment of demographers.) Similarly, the death rate of 30-year-old women is the number, per thousand women aged 30 at the beginning of a year, who die during the year. The use of one thousand women of each age as a reference group is a convenience to avoid decimal points.

Now suppose the age-specific birth rates and death rates were constant over time. In considering the possibility that the birth and death rates may be constant over time, we implicitly assume that males are present in sufficient numbers at the right ages to render possible the supposed fertility of the female population. We also assume that the birth and death rates are not themselves functions of the size of the female population in any or all age groups (technically, that birth and death rates are not density-dependent, at least over the time horizon of interest).

Whatever rates we used to relate this year's census to last year's census, we use them again to predict next year's census from this year's census, and then the following year's census from the census of next year, and so on into the indefinite future.

What happens to the *total size* of the female population?

What happens to the *fraction* of all females who are in age class i?

While you are thinking, and I hope, guessing, let me point out that the mere posing of these questions reveals the discovery that they have regular and general answers. The answers to both questions are of practical use to demographers in developing countries. Far more impressive to me is the mathematical and scientific discovery that these questions have *simple* answers that are not interminable enumerations of special behaviors under varying conditions.

What happens to total female population size and the fraction of the female population in each age class is described by the strong ergodic theorem of demography—another twentieth-century discovery, whose demographic content is due mainly to Alfred J. Lotka.

Asymptotically, that is, after a long time, the female population size will change geometrically or exponentially. The exponential rate of change in female population size may be positive, so that the population increases; zero, so that in the long run it is constant; or negative, so that the population declines exponentially.

Whether the rate of change in population size is positive, zero, or negative depends entirely on the birth and death rates, and not at all on the distribution of females among the age classes when the process of projection is begun (provided there are at least some females able to give birth in the initial population, and also under still weaker conditions). The long-run rate of population change depends only on the rates of birth and death, not the initial population census.

The fraction of females in age class i, that is, the ratio of $y_i^{(t)}$ to the total female population size in year t, behaves as illustrated in Figure 10. There two different populations are both projected forward in time using a hypothetical set of constant birth and death rates. After 100 years, the population pyramids, which consist of the fractions of females in each age class, are indis-

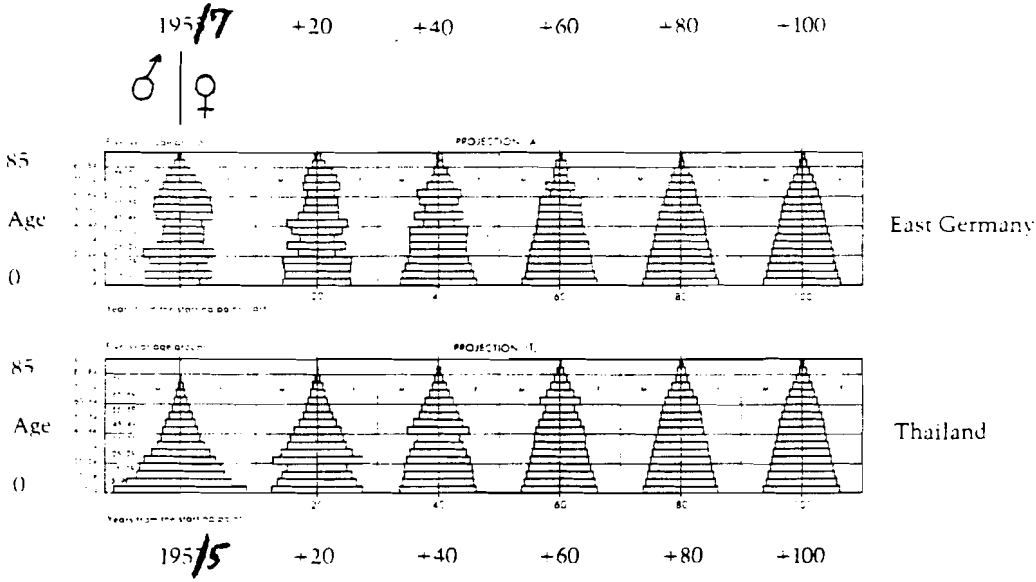

FIGURE 10
Illustration of the strong ergodic theorem of demography: If two initially different age structures experience identical birth rates and survival coefficients for a long enough time, the differences between the age structures will disappear. The upper age structure on the left describes East Germany in 1957; the lower age structure on the left estimates that of Thailand in 1955. The figure is from J. Bourgeois-Pichat, *The concept of a stable population: application to the study of populations of countries with incomplete statistics* (New York: United Nations, 1968), p. 6.

tinguishable. Figure 10 illustrates the general law that, assuming constant birth and death rates, the long-run fraction in each age class depends only on those rates and not on the distribution of females among age classes, or age structure, in the initial census. The long-run fractions of females in each age class are known collectively as the stable age structure determined by the corresponding birth and death rates.

The two main phenomena described by the strong ergodic theorem of demography, under the assumption of fixed birth and death rates, are that the population changes size asymptotically at an exponential rate that depends only on the rates and not on the initial census; and that the population asymptotically approaches an age structure that also depends only on the rates and not on the initial census.

Both phenomena are summarized by saying that populations forget their past. This forgetting makes a science of population growth rates and age structures possible. If it were necessary to know the exact census of England at the Battle of Hastings to explain quantitatively its age structure and growth rate now, demography would be a hopeless subject. (There are those who argue it is anyway. Perhaps it is, but a need for the historically unknowable is not one of the reasons why.)

Review and Conclusion

I have now given you an aerial reconnaissance of three ways in which the past may be related to the future: the coin-tossing model, as a model of

independent choices; the urn model, as a model of dependent choices; and the stable population model, as a model of deterministic interactions.

Consider now some general questions about how the past is related to the future in the light of these examples.

Do early events matter greatly? Yes, for coin tossing, if you are interested in who takes and keeps the lead. Yes, for the urn model. No, not at all, for the population model: growth rates and age structures are independent of early history.

Do long stretches of all replications or simulations of the process look alike? Yes, for coin tossing, because most of the time one player has the lead and keeps it; changes of lead are rare. No, for the urn model: the asymptotic fraction of A balls varies randomly from one simulation to the next. Yes, for the population model: any long stretch of time will reveal population size changing exponentially and constant age structure.

Does the process revisit its initial conditions? Yes, for coin tossing: the two players return to being exactly tied infinitely often. No, for the urn model in general; the proportion of A balls in the long run will be exactly ½ with negligible probability. No, for the population model, unless the initial age structure just happens to be the stable age structure.

Finally, are practical uses of the model important in the real world? Yes, yes, and yes! The coin-tossing model provides the underlying theory for nonparametric statistical tests of goodness of fit that are widely used in applied, including industrial, statistics. Models with a formal structure very similar to that of the urn model arise in population genetics and are used in the planning and analysis of pig breeding. The theory of stable age structures has found wide use in estimating the demographic characteristics of countries with incomplete demographic data.

Since none of the three examples has a monopoly on practical usefulness, and since the other characteristics of the three examples are so widely divergent, I am left, and I leave you, with the problem of deciding: Which of these examples, if any, is the world really like?

It is as if a talk entitled "What Is Life?" described a bacterium, a redwood tree, and an ant colony. The question would remain unanswered, but the examples might give you an idea of what some of life's possiblities are.

The question "How is the past related to the future?" remains unanswered here. One reason for our difficulty in answering it is that we are not especially good at inferring the elementary process by which a system changes from one instant to the next, given observations of a system for a long time (e.g., the lead in coin tossing or a single realization of the urn model). Apart from such toy examples, a major part of the problem is that the universe is a live performance that is being given only once. We cannot replay the universe, or even any large chunk of it, under the same initial conditions to see what would happen on a second try. Replication is often the key to analysis, and replication on the scale germane to human and natural history is difficult.

A second reason for our difficulty is that, given the elementary process by which a system changes from one instant to the next, we are not especially good at inferring what the long-run behavior of the system will be (e.g., the fraction of A balls in the urn model or the approach to a stable age structure in the population model). Again apart from such toy examples, which are susceptible to complete mathematical analysis, a major part of the problem is that our brains and computers belong to, and are actually rather small parts of, the universe they aim to understand. Is it reasonable to hope that they can compute the behavior of systems that include themselves, even if they were given (which they are not) the basic laws?

The task of understanding how past and future relate in *small* pieces of the world is not hopeless, merely difficult. Globally, I do not know how the past is related to the future, and I think it unlikely that anyone else can know with ultimate certainty. But the skepticism my three examples argue for can and should be applied to my conclusions as readily as to anyone else's.