11

Mises redux

Once one has clarified the concept of random sequence, one can define the probability of an event as the limit of the relative frequency with which this event occurs in the random sequence. This concept of probability then has a well defined physical interpretation. (Schnorr, 1971, pp. 8–9)

Mises' (1919) concept of *irregular* ("random") *sequence* resisted precise mathematical definition for over four decades. (See Martin-Löf, 1970, for some details.) This circumstance led many to see the difficulty of defining "irregular" as *the* obstacle to success of Mises' program, and to suppose that the solution of that difficulty in recent years has finally set probability theory on the sure path of a science along lines that Mises had envisaged. To the contrary, I shall argue that since stochastic processes do not go on forever, Mises' identification of each such process with *the infinite sequence of outputs it would produce if it ran forever* is a metaphysical conceit that provides no physical interpretation of probability.

1. BERNOULLI TRIALS

Martin-Löf (1966) showed how to overcome the distracting technical obstacle to Mises' program, and Schnorr (1971) and others have continued his work. The air is clear for examination of the substantive claim that probabilities can be interpreted in physical terms as limiting relative frequencies of attributes in particular infinite sequences of events.

The simplest examples are provided by binary stochastic processes such as coin-tossing. Here, Mises conceives of an unknown member, h, of the set of all functions from the positive integers to the set $\{0, 1\}$ as representing *the* sequence of outputs that the process

First published by R. Jeffrey, in *Basic Problems in Methodology and Linguistics*, R. E. Butts and J. Hintikka, eds., 1977. Reprinted by permission of Kluwer Academic Publishers.

would produce if it ran forever. He then identifies the physical probabilities of attributes as the limiting relative frequencies of those attributes in that sequence; e.g., in the case of tosses of a particular coin, h is defined by the condition

(1) $h(i) = 1$ *iff the ith toss (if there were one) would yield a head,*

and the probability of the attribute *head* is defined,

$$(2) \qquad p(head) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(i).$$

Both parts of this definition are essential to Mises' attempt to interpret $p(head)$ as a physical magnitude.

In their algorithmic theory of randomness for infinite sequences, Martin-Löf, Schnorr, et al. have provided satisfactory abstract models within which part (2) of the definition makes mathematical sense. Thus, Martin-Löf (1968) proposes a model in which Mises' irregular collectives are represented by the set of all functions h that belong to all sets of Lebesgue measure 1 that are definable in the constructive infinitary propositional calculus, e.g., the set of sequences for which $p(\text{head}) = 1/2$ in (2). In proving that the intersection of all such sets has measure 1, he shows that his definition escapes the fate of von Mises' (according to which there would be no random sequences) and yields the desired result, that "almost" all infinite binary sequences are random. The condition $p(head) = 1/2$ is inessential: The same approach works for Bernoulli trials with any probability of *head* on each.

But the brilliance of this abstract model of Bernoulli trials is far from showing how probability is connected with physical reality: Rather, it deepens the obscurity of Mises' condition (1), which purports to provide that connection. For most coins are never tossed, and those that are, are never tossed more than finite numbers of times. No infinite sequence of physical events determines the function h of (2): For all but a finite number of values of "i," the clause following "iff" in (1) must be taken quite seriously as a counterfactual conditional. But unless the coin has two heads or two tails (or the process is otherwise rigged), there is no telling whether the coin would have landed head up on a toss that never takes place. That's what probability is all about.

A coin is tossed 20 times in its entire career. Would it have landed

head up if it had been tossed once more? We tend to feel that there must be a truth of the matter, which could have been ascertained by performing a simple physical experiment, viz., toss the coin once more and see how it lands. But there is no truth of the matter if there is no 21st toss. The impression that there *is* a truth of the matter arises through the analogy between (a) extending a series of tosses of a coin, and (b) extending a series of measurements of a physical parameter, e.g., mass of a certain planet. If *p(head)* is a physical parameter on a par with *m*(Neptune), then – the argument goes – (a) really is just like (b). But the analogy is a false one because while Neptune exists, and has a mass whether or not we measure it to a certain accuracy, the 21st toss of a coin that is tossed only 20 times does not exist and has no outcome: Neither *head* nor *tail*. A truer analogy would compare *p*(head) with *m*(*x*) where *x* is a nonexistent planet, e.g., the 10th from the Sun. Mises defines *p(head)* as the limiting relative frequency of heads in an infinite sequence that has no physical existence. If one could and did toss the coin forever (without changing its physical characteristics) one would have brought such a sequence into physical existence, just as one would have brought an extra planet into existence by suitable godlike feats, if one were capable of them and carried them out. But in the real world, neither the sequence nor the planet exists, and the one is as far from having a limiting relative frequency of heads as the other is from having a mass.

Granted: There is a telling difference between the two cases. In the case of the nonexistent 10th planet we are at a loss to say what its mass would be if it had one, while in the case of the coin that is tossed just 20 times we are ready enough to name a probability for heads. If the coin is a short cylinder with differently marked ends and homogeneous mass distribution, we are confident that heads have probability 1/2. But this difference tells against Mises: It identifies the probability of heads as a physical parameter of the coin, whether or not it is ever tossed, in terms of which we explain and predict actual finite sequences of events – directly, and not by reference to a nonexistent infinite sequence of tosses. It is because the probability of heads is 1/2 that we grant: If the coin *were* tossed ad infinitum without changing its physical characteristics, the limiting relative frequency of heads would be 1/2. But since there is no

infinite sequence of tosses, "its" characteristics cannot explain why heads have probability 1/2.

2. IRREGULAR FINITE SEQUENCES

In the 1960s, Kolmogorov and others (Chaitin, Solomonoff) founded a theory of algorithmic complexity of finite sequences that sheds fresh light on probability. In showing that the sequences irregular in Kolmogorov's sense are those that pass a certain universal test for randomness, Martin-Löf (1966) provided an alternative definition of irregularity that he was able to extend quite naturally to the case of infinite sequences. In deprecating the foundational importance of the infinite case, I am far from denying the importance and foundational relevance of the finite case as treated by Kolmogorov, Martin-Löf, and others. What I do wish to deny is that by continuity with the finite case, or by mathematical infection from it, the infinite case gets the importance it would have if ours were a world in which each Bernoulli process went on forever (and in which each Markov process, infinitely replicated, went on forever). To get a sense of the importance and autonomy of the finite case, let us review it briefly.

Tables of "random numbers" are long, irregular sequences of digits – binary digits, let us suppose. The easiest and surest way to generate such sequences is by Bernoulli processes with equiprobability for the two outcomes on each trial, e.g., by repeated tosses of a coin, with heads recorded as 1's and tails as 0's. In principle, such a process could yield a table of a million 1's, but in practice, no one would buy such a table or give it shelf space.

Why? Well, why spend the money? The table is utterly regular, the relevant rule being, "Write 1 000 000 ones." It is not only cheaper but easier to use that rule in your head than to buy and consult the table. *Moral:* We use "equi-Bernoulli" processes to generate tables of "random numbers" not because we have use for the outputs of such processes no matter what they may prove to be, but because we expect such outputs to be irregular, and it is irregularity of the sequence that we seek, irrespective of its provenance.

Kolmogorov (1962) pointed to *incompressibility* as a definitive characteristic of irregularity of finite sequences. Thus, a string of 1 000 000 1's is compressible, for the rule "Write 1 000 000 ones"

would be only some 100 binary digits long if letters, digits, and spaces were coded in some fairly simple way as blocks of binary digits. In detail, questions of compressibility are relative to (1) choice of one out of the infinity of universal systems of algorithms or programming schemes for generating binary sequences, and to (2) choice of one out of the infinity of measures of complexity of algorithms belonging to the same universal system – let us say, via length of representation in one of the infinity of effective binary coding schemes. Once these choices have been made, we have the means to define the *irregular* ("random") finite sequences as those *about as long as the shortest binary coded algorithms that generate them.* If we define the *algorithmic complexity* of a finite binary sequence as *the length of the shortest binary coded algorithms that generate it,* then the irregular sequences are those whose lengths are approximately equal to their algorithmic complexities.

Locally (i.e., for each particular sequence) the relativity of algorithmic complexity to choices (1) and (2) is problematical [cf. Goodman's (1955) "grue" paradox], but globally its effect is negligible, for if k_1 and k_2 are two particular measures of algorithmic complexity, there will be a finite bound on the absolute differences between $k_1(s)$ and $k_2(s)$ as "s" ranges over all finite binary sequences. Thus, one proves that the percentage of irregular sequences among all sequences of the same length approaches 100 as the length of the sequences increases without bound: *For large n, practically all sequences of length n are irregular.*

Why do we turn to equi-Bernoulli processes as sources of irregular finite sequences? Kolmogorov's theory provides a clear answer, as follows. (1) For such processes, all output sequences of length n have probability 2^{-n}. (2) For large n, practically all sequences of length n are irregular. Therefore: (3) For large n, the probability is practically 1 that the output of such a process will be irregular. Then devices lie ready to hand that, with practical certainty, generate long irregular sequences. But mathematical certainty about irregularity is far more difficult to attain: (4) For universal systems of algorithms, the halting problem is unsolvable, and therefore there is no effective test for irregularity of finite sequences. In principle, one might nevertheless be able to prove that particular finite sequences are irregular, but in practice we do well to rest content with high probability.

Suppose that a coin is tossed 40 times, and the process yields nothing but heads. There is a dim argument to the effect that this should not surprise us, for the sequence of 40 heads is no less probable than any other sequence of that length, be it ever so irregular. Of course, this argument must be wrong if we rightly see in such an output compelling evidence that the source was not as we had supposed it to be, e.g., if we see the output as overwhelming evidence that the source, far from being equi-Bernoullian, is one that yields heads with probability 1 on each toss. But what is the rationale behind this sensible view of the evidence? Here I give a mixed Bayesian answer to this question – "mixed" in the sense that while statistical hypotheses about the source are treated objectivistically (as hypotheses about physical magnitudes), probabilities of those hypotheses are treated judgmentalistically ("subjectivistically").

(1) Consciously or not, we do or should entertain various hypotheses $H_1, H_2, \ldots$ about the source, where (in the present example) initially we judge H_1 (the equi-Bernoullian hypothesis) to be overwhelmingly more probable than H_2 (the hypothesis that heads have probability 1 on each toss), in some such sense as this:

$$\frac{p(H_1)}{p(H_2)} = 2^{20} \approx 1\ 000\ 000.$$

(2) After seeing the output sequence and so verifying the evidence-statement E = "The output is a string of 40 heads," we revise our judgment via the probability calculus, changing our degree of belief in each hypothesis H from its prior value, $p(H)$, to its posterior value,

$$p(H \mid E) = p(E \mid H) \times \frac{p(H)}{p(E)} \quad \text{(Bayes' Theorem).}$$

so that now H_1 is overwhelmingly less probable than H_2, in the sense that

$$\frac{p(H_1 \mid E)}{p(H_2 \mid E)} = \frac{p(E \mid H_1)p(H_1)}{p(E \mid H_2)p(H_2)} = 2^{20}2^{-40} \approx 0.000\ 001.$$

In this Bayesian answer, the probabilities of the hypotheses are "subjective" in the sense that they are degrees of belief, which need

not be "subjective" in the sense of being ill-founded, arbitrary, or idiosyncratic. But the statistical hypotheses H_1 and H_2 themselves are treated objectivistically. Some Bayesians – notably, de Finetti – would treat all probabilities as degrees of belief, and others would treat all of them objectivistically. The mixed position represented here is the commonsensical version of Bayesianism that Bayesian extremists must explain away or reproduce within their own terms of reference.

"Bayesians" are so called because of their willingness to use Bayes' theorem in cases where most thoroughgoing objectivists would reject as senseless the prior probabilities $p(H)$ and $p(E)$ of evidence and hypothesis that appear in it. The affinity of Bayesianism with "subjectivism" (judgmentalism) derives from the fact that we may have broadly shared judgments in the form of degrees of belief in H and E even in cases like the present example, where prior inspection of the coin is supposed to have led us to think the equi-Bernoullian hypothesis overwhelmingly more probable than the other, but where we envisage no definite stochastic process of which the coin is the product – a process of which the ratio of *physical* probabilities would be $p(H_1)/p(H_2) \approx 1\ 000\ 000$. Pure objectivists who would be Bayesian must envisage some such higher-level process, and treat the prior probability function p as the probability law of that process. Thus, commonsense objectivists sometimes speak (without conviction) of urns containing assortments of coins, some normal, some bent, some two-headed, etc., out of one of which the coin actually used is imagined to have been drawn. In that vision, $p(H_1)$ is the proportion of normal coins in the urn.

Observe that where the "subject" thinks she knows the objective probability of an event (e.g., the event that all 40 tosses yield heads) and thinks she knows nothing else that bears on the matter, (e.g., perhaps, that the first toss yielded a tail!), she will adopt what she takes to be the objective probability as her degree of belief. Then "subjective" does not mean *whimsical*. To call a probability "subjective" is simply to say that it is somebody's degree of belief. One does not thereby deny that the belief has a sound objective basis. Furthermore, the "events" to which subjective probabilities can be attributed need have no special character (e.g., "unique," weird, etc.), for they are simply the events concerning which people can

have degrees of belief, viz., all events whatever. [These remarks are directed in part to the comments on subjective probability in Schnorr (1971, p. 10).]

4. NONFREQUENTIST OBJECTIVISM

Frequencies are important: The laws of large numbers tell us why; e.g., they tell us that in stationary binary processes, the relative frequency of "success" will in all probability be very close to the probability of success on the separate trials. Notice that here, the notion of probability appears along with that of relative frequency in the formulation of the law itself. (The notion of probability appears as well in the definition of "stationary," viz., invariance of probability of specified outcomes on specified trials, under translation of trials.) Frequentism is a doomed attempt to define probability in such a way as to turn the laws of large numbers into tautologies.

The lure of von Mises' program lies in its goal of providing a uniform, general definition of probability as a physical parameter – a definition that can be applied prior to the scientific discoveries that reveal the detailed physical determinants of stochastic processes, as e.g., the discoveries by Mendel, Crick, and many others revealing the mechanisms underlying the mass phenomena encountered in genetics. Mises sought to found probability as an independent science, on the basis of imaginary infinite sequences of events. Taxed with the unreality of those foundations, he replied that they are as real as the foundations of physics: To measure the physical parameter prob(*head*) to a desired accuracy it suffices to toss the coin often enough, for prob(*head*) is the limit of such a sequence of measurements just as surely as m(Neptune) is the limit of another sequence of measurements. Shall we hold the foundations of probability to a higher standard of physical reality than that to which we hold physics itself?

Surely not; but here, Mises holds physics itself to a remarkably low standard of reality, i.e., essentially, the idealist standard to which Bishop Berkeley held it: *Esse percipi est*. The suggestion is that the mass of Neptune exists to the extent to which we measure it, just as the sequence of outcomes exists to the extent to which we toss the coin. As was suggested in Section 1, the limiting relative

frequency of heads in the "ideal" (i.e., nonexistent) infinite sequence of tosses is more properly compared with the mass of some nonexistent planet, e.g., the 10th from the Sun.

But if probabilities are not limiting relative frequencies, what are they? If there is no uniform, general definition of probability that is independent of other scientific inquiries, how shall we define probability as an objective magnitude? I would answer these questions as follows.

The physical determinants of probabilities will vary from class to class of cases; there is no telling a priori what they will prove to be. In the case of die-casting, the experience of gamblers and tricksters joins with physical and physiological theory to point to the shape, mass distribution, and (most important) markings of the die itself, as the determinants of the probabilities of the possible outcomes on each toss, and these considerations also join to say that different tosses are probabilistically independent. The case is similar for coin-tossing (where the point about markings is that there *are* two-headed coins about). In lotteries, by design, the determinants are the numbers of tickets of each sort (or the numbers of balls of different colors in the urn), but design is not enough: Empirical and theoretical inquiry may show the design to have been defective, e.g., because the balls of one color share a palpably distinct texture. As with games of chance, so with social, biological, and physical probabilities, but even more so: We look to experience, informed with theory, to identify the objective determinants of the probability laws of types of stochastic processes.

The easiest cases are lotteries and urn processes. There, we identify objective statistical hypotheses with the makeup of (say) the urn, and, by a happy accident, the probability of drawing a ball of a certain color is numerically equal to the proportion of balls of that color in the urn. In practically all other cases, such a numerical coincidence is lacking. The "classical" view tried to generalize that coincidence to all stochastic processes. The frequentist view tries to generalize a different coincidence – one that is probable where the law of large numbers holds. On a nonfrequentist objectivistic view, one must face the fact that typically, no such coincidence will be forthcoming – not uniformly in all cases, and not even differentially, on a case-by-case basis. Still, we are often in a position where we can be fairly sure that the relevant determinants, difficult as they

may be to describe explicitly and in detail, are the same in two processes, as when we ascertain that two coins were cast in the same mold under similar conditions: Believing that the determinants are shape, mass distribution, and markings, and having good reason to think that these determinants were determined in the same way for the two coins, we have good reason to think that the same probability law will govern the two processes of tossing them – even though we are at a loss to specify the common shape or the common mass distribution except ostensively.

No pure objectivist, I think it important to use judgmental probabilities, e.g., as illustrated in §3 (in an extreme, simplified example). The present suggestion is that the objective statistical hypotheses to which judgmental probabilities are attributed in such cases will be hypotheses about various kinds of physical magnitudes, which we shall seldom be in a position to specify explicitly and in detail, but which we can often identify ostensively, well enough for our purposes, once we understand what the *kinds* of magnitudes are that determine the process at hand – kinds like shape, mass distribution, and marking.

This is a far cry from Mises' uniform, general identification of probability with a particular physical magnitude, found in all cases; but that magnitude does not exist.

REFERENCES

Chaitin, G. J.: 1966, "On the Length of Programs for Computing Finite Binary Sequences," *J. Assn. Computing Machinery* **13,** 547–569.

Chaitin, G. J.: 1969, "On the Length of Programs for Computing Finite Binary Sequences: Statistical Considerations," *J. Assn. Computing Machinery* **16,** 145–159.

DeFinetti, B.: 1974, 1975, *Theory of Probability,* Wiley, 2 vols.

Goodman, N.: 1955, *Fact, Fiction, and Forecast,* Harvard.

Kolmogorov, A. N.: 1963, "On Tables of Random Numbers," *Sankhyā,* Ser. A **25,** 369–376.

Kolmogorov, A. N.: 1965, "Three Approaches to Definition of the Concept of Information Content" (Russian), *Probl. Peredači Inform.* **1,** 3–11.

Martin-Löf, P.: 1966, "The Definition of Random Sequences," *Information and Control* **6,** 602–619.

Martin-Löf, P.: 1970, "On the Notion of Randomness" in A. Kino et al. (eds.), *Intuitionism and Proof Theory* (Proc. of Summer Conf., Buffalo, N.Y., 1968), North-Holland.

Mises, R. v.: 1919, "Grundlagen der Wahrscheinlichkeitstheorie," *Math. Z.* **5**, 52–99.
Mises, R. v.: 1928, 1951, *Probability, Statistics, and Truth* (2nd revised English ed.), Macmillan, 1957.
Mises, R. v.: 1964, *Mathematical Theory of Probability and Statistics*. Academic Press.
Schnorr, C. P.: 1971, *Zufälligkeit und Wahrscheinlichkeit*, Springer Lecture Notes in Mathematics **218.**
Solomonoff, R. J.: 1964, "A Formal Theory of Inductive Inference," *Information and Control* **7**, 1–22.

12

Statistical explanation vs. statistical inference

Hempel is not the first philosopher to have held that causal explanations are deductive inferences of a special sort. In the *Posterior Analytics*[1] Aristotle distinguishes a special sort of deductive inference – the demonstrative syllogism – in these terms:

> By demonstration I mean a syllogism productive of scientific knowledge, a syllogism, that is, the grasp of which is *eo ipso* such knowledge.

He then lays down defining conditions for this special sort of inference:

> . . . the premisses of demonstrated knowledge must be true, primary, immediate, better known than and prior to the conclusion, which is further related to them as effect to cause.

And he remarks,

> Syllogism there may indeed be without these conditions, but such syllogism, not being productive of scientific knowledge, will not be demonstration.

Now we can fault this account on various grounds, but so can we fault contemporary accounts. We must give the old man credit; as he says at the end of the *Organon* (at the end of *De Sophisticis Elenchis*), his was the first book on logic; and he concludes,

> . . . there must remain for all of you, or for our students, the task of extending us your pardon for the shortcomings of the inquiry, and for the discoveries thereof your warm thanks.

The affinities between the Hempelian and Aristotelian accounts of explanation may be obscured by differences in terminology. Thus, Aristotle speaks of syllogism, Hempel of deductive inference; and Aristotle speaks of knowledge, Hempel of explanation. But remember that "syllogism" was Aristotle's general term for deductive

First published by R. Jeffrey, in *Essays in Honor of Carl G. Hempel*, N. Rescher et al., ed., 1969. Reprinted by permission of Kluwer Academic Publishers.

1. Book 1, ch. 2. All citations from this work are from the Oxford translation.