Probabilistic Causality, Randomization and Mixtures

Jan von Plato

University of Helsinki

1. Basic Notions

The scheme of abstract dynamical systems will represent repetitive experimentation: There is a basic space of events X' and the denumerable product $X = X' \times X' \times X' \times \ldots$ contains all possible sequences of events $x = (x_1, x_2, \ldots)$. There are projections q_n which give the nth member of x: $q_n(x) = x_n$. A transformation T is defined over X by the equation $q_n(Tx) = q_{n+1}(x)$. It removes the sequence by one step, $T(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$ and is known as the shift transformation. It comes as an abstraction of the dynamical transformations of classical theories. Here it represents the performance of 'the next' experiment. There is a field of subsets $A \subset X$ and $TA = \{Tx | x \in A\}$, $T^{-1}A = \{x | Tx \in A\}$. A probability measure P over X is *stationary* if $P(T^{-1}A) = P(A)$. If one experiment is performed at each unit interval of time, the probability of the event A is "now" the same as the probability of the event $T^{-1}A$ was in the preceding experiment. It follows that $P(\{x | x_{n+1} = i_1, \ldots, x_{n+k} = i_k\})$ is independent of n, so that the usual expression for stationarity as invariance of probabilities in time is recovered. Sets for which $T^{-1}A = A$ are called invariant. A stationary measure P is *ergodic* if invariant sets have measure zero or one.

Let us now suppose that we have simple events; each $x_i = 0$ or 1 with $x_i = 1$ as "occurrence of the event at trial i". The central result of the theory states for this simple situation that the limit of relative frequency $\lim_{n\to\infty} \Sigma x_i / n$ of the event exists for almost all sequences x. The exceptions have P-measure zero. If P is ergodic the limit is the same for almost all sequences. Both results hold also in the other direction, so that stationarity is equivalent to the existence of limits of relative frequencies, and ergodicity to their uniqueness. Since $x_i = q_1(T^i x)$, with T^i the i-fold iteration of T, the two results can be ex-

PSA 1986, Volume 1, pp. 432-437

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press

pressed in another terminology as follows: The *time average* $\lim_{n\to\infty} \sum_i q_1(T^i x)/n$ of the function q_1 along the *trajectory* x exists under stationarity, and is unique (independent of x) under ergodicity. The same holds for any measurable function of x. The terminology comes from the theory of classical dynamical systems where the state space X' is continuous and T is substituted by a group of mappings corresponding to solutions of the equations of motion. Ergodicity characterizes the case where it is possible to asymptotically identify probabilities as limiting frequencies along one sequence of repetitions.

2. Probabilistic Causality

A function g over X is an *invariant of motion* if its value is preserved under the dynamics T: $g(x) = g(Tx) = g(T^i x)$ ... Equivalence classes $g^{-1}[y] = \{x|g(x)=y\}$ having the same value of g, partition the space X. If g has finitely many values, there are finitely many parts.

Next we note that time averages are invariants of stationary systems. Intuitively, the limit over $x = (x_1,x_2,...)$ does not depend on finite segments $(x_1,...,x_k)$. The time average $\hat{f}(x)$ of f over x is $\lim_{n\to\infty} \sum_i f(T^i x)/n = \lim_{n\to\infty} \sum_i f(T^{k+i}x)/n = \hat{f}(T^k x)$ so that $\hat{f}(x)$ is an invariant of motion. (Proofs of results referred to may be found in, e.g., Cornfeld, Fomin and Sinai 1982.)

The indicator function I_X of the whole space has a constant value 1 over X and is therefore an invariant. Invariant functions partition the space into classes which are invariant sets as can be seen from above definitions. If one of these sets has P-measure strictly between zero and one, ergodicity fails. In the other direction, the indicator of an invariant set is an invariant function. This amounts to the result that ergodicity is equivalent to having essentially only one invariant function. Its value must be constant over state space, and other possible invariant functions are determined by the value of the independent invariant. The total energy is an example for classical dynamical systems. Its counterpart in the abstract setting is the (not necessarily uniform) ergodic P-measure over X. Time averages of indicator functions I_A are invariant. For ergodic systems these averages $\hat{I}_A$ equal the P-measure P(A). For the classical case, it follows that probabilities are determined by total energy.

Let us next suppose that ergodicity fails for P. Further, let there for the sake of simplicity be only one additional independent invariant of motion f with a finite range of values $y_1,...,y_k$. The components $f^{-1}[y_i] = A_i$ of the corresponding partition of X are invariant sets. We assume further that they each have positive measure $P(A_i)$. As there are no further invariants, each of the A_i determines a measure $P_i(B)=P(B|A_i)$ over X. It is ergodic over A_i. Writing $P(A_i) = a_i$, total probability

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press

gives the representation $P(B) = \sum_i a_i P_i(B)$. Our simplifying assumptions have led to a special case of a result which in general terms states that every stationary measure has a unique decomposition into ergodic parts (see my 1982).

A *causal factor* is for us something with which a difference can be made in the performance of an experiment. It may be the case that experiments can be so prepared that there is only one possible result. For classical dynamical systems, this would be accomplished by finding as many independent invariants of motion as there are degrees of freedom in the system under consideration, let us say n. Now, fixing n values of suitable quantities as initial conditions, the result is unique. In general, this would not be the case. The system has a high n (i.e., a complete description is very complex), whereas the number of independent invariants k is much less than n. We have seen that if the experimental arrangement remains unaltered, stationarity may be assumed. Repetitive experimentation leads in this way to stable statistical behaviour, at least asymptotically if not faster. Only, it varies from trajectory to trajectory. A determination of that behaviour is only possible, as a prediction from theory and initial data, if it is known what values obtain for the invariants of motion. That is, if we have a complete set of invariants $f_1,\ldots,f_k$, with fixed values, knowledge of the trajectory x is not needed for the calculation of statistical laws. More generally, all properties of a system are identified as functions f over X. If f is invariant, its value f(x) is determined from $f_1(x),\ldots,f_k(x)$ which are complete in exactly this sense. In the other direction, by controlling the values of a complete set of invariants, we are able to determine and control the statistical laws that obtain for our experimental arrangement. Therefore, *probabilistic causality* - causing events to occur with a given statistical law - requires the identification of the relevant *causal factors as invariants of motion* that form a complete set. This is, of course, a highly ideal result. It tells how statistical analysis would proceed if we were as competent in general experimental situations as we are in a handful of examples from physics. We would determine the permanent, invariant features of the arrangement, and the dynamics would single out a statistical law which is preserved under the dynamics. No additional gathering of data would be needed. In applications, we would prepare the system so that it produces those of its possible statistical patterns we most want it to produce. Incidentally, the above also gives us a notion of probabilistic explanation in the sense of explanation of a probabilistic law, from values of invariants obtaining in the situation and from the dynamical law.

3. Randomization and Mixtures

If $\{P_s\}_{s\in S}$ is a family of probability measures over X, any convex combination $\sum_i a_i P_i$ with $a_i \geq 0$ and $\sum_i a_i = 1$ and each P_i in $\{P_s\}_{s\in S}$ is again a probability measure over X. In the most general formulation there is a measure F over the arbitrary index set S, and *mixtures* have the form $P(A) = \int_S P_s(A)dF$. Feller (1971, pp. 53 - 58) has suggested that forming a mixture over a suitable parameter may be described probabilistically as randomization. Our above notion of causal factors is applicable

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press

here. For simplicity, let us assume that there is a parameter $\lambda \epsilon \Lambda$ such that each P_λ is ergodic, and a measure m over Λ, so that mixtures are of the form $\int_\Lambda P_\lambda(A)dm(\lambda)$. Obviously, λ takes here over the task of invariants of motion.

Suppose now that we fix the value of λ. P_λ is a statistical law of the component of X corresponding to the given λ. If the value of λ is fixed but unknown, we will be able to identify P_λ asymptotically from data. It will be instructive to think of a slightly different situation in which the events occur also simultaneously, not only sequentially. Instead of one "test item", we have a great number of them tested simultaneously. This will form a "population". It is investigated how this population can be *stratified* according to its statistical properties. In repetition, each item gives a sequence of results: $x_1=(x_{1_1},x_{1_2},\ldots)$, $x_2=(x_{2_1},x_{2_2},\ldots),\ldots$ Conditions prevailing for test item number i are summarized in some value λ_i of λ. Now, if it is our aim to obtain statistical data from the whole population, λ must be varied from experiment to experiment. Moreover, each value of λ has to be weighted by the proportion of test items that are characterized by that value of λ. The result is statistical data from the whole population, *randomized* with respect to properties between which λ discriminates. In an ideal situation, it is known what causal factors λ there are. Randomization of experiments will be perfect.

Let us first distinguish between randomized and "one-component" repetitions. In the former, the value of λ is chosen according to m each time a repetition is made. The effect is that the statistical properties of the population as a whole are identifiable through time averages. Here the population and variation of λ are real. In the second case, one randomization is performed, after which the chosen component remains. It follows that only that component's P_λ is identified. Statistical properties of other components do not become manifest. A second distinction is made between cases in which the initial randomization is real and those in which it is fictive.

Let us now stipulate that *causal independence* is violated if the performance of a repetitive experiment changes the experimental arrangement of further experiments. Simple examples are offered by drawings from finite urns without replacement. A contrary case certainly obtains if events with space-like separation are considered.

According to above distinctions, there are at least three different ways of tossing bent coins. Everything is supposed discrete for simplicity. Each bent coin has some probability λ_i of landing heads. This is represented by an urn of black and white balls with proportion λ_i of white balls. Our three cases appear as follows: 1. There is a bag of urns such that each urn i with probability λ_i appears in a definite proportion a_i. Randomized experimentation consists of first drawing an urn, then drawing a ball, and replacing both. This procedure is re-

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press

peated. The probability of "heads" (with white for heads) is, from total probability, $\sum_i a_i\lambda_i$. Successive events are probabilistically independent, and the limit of time average of "heads" equals the total proportion of white balls, as if the walls of the urns went broke. 2. Only one randomization is made. If it gives urn i, probability of heads is λ_i, with independent repetitions. If the urn cannot be identified, the probability is $\sum_i a_i\lambda_i$ and successive events are probabilistically dependent. In the latter, λ_i is however asymptotically identifiable from data. 3. There is only one urn of unknown composition. Uncertainty is represented by weights a_i for the different possible compositions (ratios of white balls) λ_i. Probability of heads on first toss is $\sum_i a_i\lambda_i$ and successive events are, as above, dependent. Typically, given λ_i, one assumes independence, and this makes the mixture probabilities exchangeable. If the urn is not identified or if there is only one urn, successive events are probabilistically dependent, whereas the composition of the urn remains identical. Causal independence is *not* violated.

The reader will have noticed that de Finetti's notion of exchangeability fits well into the above. In fact, the notions of section 1. above are a general way of putting these matters, with exchangeability as a special case of stationarity, independence a special case of ergodicity, and de Finetti's representation theorem a special case of the ergodic decomposition theorem. de Finetti wants to argue that the objective probabilities (λ_i above) are fictive entities, and indeed reducible to the subjective ones (the mixtures $\Sigma a_i\lambda_i$ above). For us, the interpretation is rather the reverse: the mixtures determine a unique set of weights for the λ_i. The a_i are either weights in a randomized experiment, or else fictive entities. Probabilistic dependence reflects the fact that we have not identified the limit of time average λ_i.

4. Note on Applications in Physics

Under another nomenclature, our causal factor λ above appears as a hidden variable in foundations of quantum theory. If the distribution of λ is ergodic, there cannot be any causal factors that λ would bear. Such a theory would simply reproduce the statistical laws of quantum theory. Therefore, there should exist functions invariant with respect to P_λ, that is, causal factors affecting the statistics of quantum experiments. (Such an approach was attempted in the early sixties by Daneri, Loinger and Prosperi as is well known.) However, the curious correlations of some quantum experiments cannot be explained as effects of such causal factors. This is excluded since even an ergodic P_λ is not permitted by Bell's inequality, much less mixtures that would factorize the ergodic measure.

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press

References

Cornfeld, I.P.; Fomin, S.V.; and Sinai, Ya.G. (1982). Ergodic Theory. Berlin: Springer.

Daneri, A.; Loinger, A.; and Prosperi, G.M. (1963). "Quantum Theory of Measurement and Ergodicity Conditions." Nuclear Physics 33: 297-319.

de Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives." Annales de l'Institut Henri Poincaré 7: 1-68.

-------------. (1938). "Sur la condition d'équivalence partielle." Actualités Scientifiques et Industrielles 737: 5-18.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications. Volume II. 2nd ed. New York: Wiley.

von Plato, J. (1982). "The significance of the ergodic decomposition of stationary measures for the interpretation of probability." Synthese 53: 419-432.

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193143 Published online by Cambridge University Press