TESTING PROBABILISTIC CAUSALITY

ABSTRACT. Probabilistic causality is related to the value of $\lambda$, the parameter of the predictive probability function elaborated by Gini, Johnson, and Carnap. Thus it is impossible to make estimates within Suppes's theory of probabilistic causality. This causality can only be checked using tests of significance. Two examples of this are given. Some considerations about the corroboration of a probability law conclude the paper.

1. In his celebrated monograph A Probabilistic Theory of Causality Patrick Suppes defines quadrant causality in the following way:

The property $Y_{t'}$ is a (weak) prima facie quadrant cause of the property $X_t$ if and only if (i) $t' < t$. (ii) For all $x$ and $y$ if $P(Y_{t'} \geq y) > 0$, then $P(X_t \geq x \mid Y_{t'} \geq y) \geq P(X_t \geq x)$ (Suppes, 1970, p. 61).

Before giving this definition, Suppes points out that

I shall mainly draw on a more general and abstract theory of dependence for random variables (or quantitative properties) developed by Lehmann [...]. I use the standard notation of a random variable [...] as the basic formal tool of analysis. Each quantitative property will be represented by a random variable ... (ibid, p. 60).

It seems to me that with this assumption Suppes, following Lehmann, is dealing with populations. More precisely, the values of random variables which he considers are attributes of modalities and, consequently, the distribution of a random variable is related to the (limit of the) relative frequencies with which members of the population bear attributes. Of course, Suppes is right to maintain that “No one view of the nature of probability is required to accept the causal analysis given thus far” (ibid, p. 48). But taking random variables as representing modalities, one is speaking about relative frequencies. Hence, notwithstanding his declaration of neutrality, Suppes is introducing a strong bias in favour of a frequentistic interpretation of probabilistic causality. An example of Suppes, taken from M. G. Kendall and A. Stuart (1961, pp. 331–332) who discussed crime rate and church membership, makes what I mean clear. The five random variables considered in the example are: number
of known offenses per thousand inhabitants; number of church members of 13 years of age or over per 100 of total population of 13 years of age or over; percentage of male inhabitants; percentage of total inhabitants who are foreign-born males; number of children under 5 years old per 1000 married women between 15 and 44 years old.

But random variables can be used in another way. These variables may represent individuals too. In this case the value that a random variable assumes is the (index of the) attribute which the considered individual bears. For example \( X = 58 \) can be interpreted as affirming that Rossi Tito is 58 years old and \( \Pr\{ X = 58 \} \) as the absolute probability that Rossi Tito is in fact 58 years old. Taking this approach, and considering a modality with \( k \) attributes, a population becomes a random process

\[
(1) \quad X_1 = j_1, \quad X_2 = j_2, \ldots, \quad X_n = j_n, \ldots \quad X_N = j_n, \ldots
\]

in which the last three dots are to be considered only if the population has an infinite number of members. The indexes of the random variables of (1) are intended to denote the order in which individuals have been or will be observed.

If we are able to determine the value of

\[
(2) \quad \Pr\{ X_{n+1} = j \mid E^n \}, \quad 1 \leq j \leq k,
\]

where \( E^n := X_1 = j_1, \ldots, X_n = j_n \) is the evidence describing a sample of size \( n \) whose observation \( k \)-tuple is \( n = (n_1, \ldots, n_j, \ldots, n_k) \), \( \sum_j n_j = n \), then we can also determine the probability of (1). This holds true for any finite population. It can be proved that in some cases this holds true for infinite populations too. Hence, at least in principle, considering the probability of a random variable representing an individual we are also considering random variables representing modalities in a population.

If we intend to discuss causal relations between two modalities, for example heights and weights of human persons, we must consider two modalities and two random processes. I begin with the easiest case, that is, one modality and one random process.

2. When random variables represent individuals, we must be able to determine the value of (2) in order to ascertain probabilistic causality. In fact, if this is the case, we can determine

\[
(3) \quad \Pr\{ X_i = j \mid X_h = g, E^n \}, \quad n < h < i
\]
and

\[ \text{Pr}\{X_i = j \mid E^n\}, \quad n < i \]

and the comparison between (3) and (4) reveals whether \(X_h = g\) is a quadrant cause of \(X_i = j\).

One way to accomplish this task was discovered by C. Gini, W. E. Johnson and R. Carnap, who were pioneers in the field of predictive statistical inference. The suggestion of these authors amounts to setting down some conditions for (2), as:

**Regularity**, (2) is greater than zero;

**Symmetry (Exchangeability)**, (2) does not change upon interchanging random variables;

**Invariance**, apart from two constants, (2) only depends upon \(n_j\) and \(n\).

The first of the two constants just mentioned is

\[ p_j := \text{Pr}\{X_i = j\}, \quad i = 1, 2, \ldots, 1 \leq j \leq k, \]

i.e. the absolute (prior) probability of \(j\). The second is

\[ \eta := \frac{\text{Pr}\{X_i = j \mid X_h = g\}}{\text{Pr}\{X_i = j\}}, \quad j \neq g, \]

i.e. the *heterorelevance quotient* at zero.

It is easy to realize that, given the predictive prior distribution (5), \(\eta\) is another way of introducing correlations among individuals of the random process (1). In fact

\[ \text{Cov}(X_i = j, X_h = g) = p_j p_g (\eta - 1). \]

In other words, \(\eta\) is another way of introducing the quadrant causality between two individuals. In fact, the *selfrelevance quotient* at zero for \(j\), defined as

\[ \rho_j := \frac{\text{Pr}\{X_i = j \mid X_h = j\}}{\text{Pr}\{X_i = j\}} \]
is related to \( \eta \) through the equality
\[
\rho_j = \frac{1 - \eta(1 - p_j)}{p_j}.
\]

Hence, given the predictive prior distribution, fixing \( \eta \) is a way to fix \( \rho_j \), that is to fix the relation between \( \Pr\{X_i = j \mid X_h = j\} \) and \( \Pr\{X_i = j\} \), and this amounts to fixing quadrant causality. Time is not taken into account; it is implicitly considered by the indexes of random variables of (1).

Considering individuals, the problem of probabilistic causality is that of fixing the value of \( \eta \). But to estimate (2) in order to ascertain the existence of probabilistic causality gives rise to a vicious circle. Symmetry and invariance exclude some dependence. The first shuts out dependence upon order. So it follows that
\[
\Pr\{X_i = j \mid E^n\} = \Pr\{j \mid n\}.
\]
holds. Invariance shuts out dependence upon frequencies other than those favourable and contrary to \( j \). It follows that
\[
\Pr\{j \mid n\} = \Pr\{j \mid n_j, n\}.
\]
holds. But the conditions considered do not determine whether
\[
\Pr\{j \mid n_j + 1, n\} = \Pr\{j \mid n_j, n\}
\]
or
\[
\Pr\{j \mid n_j + 1, n\} > \Pr\{j \mid n_j, n\},
\]
holds, i.e. whether or not quadrant causality exists between two random variables of the process. As we have seen, correlation is related to the value of \( \eta \). This makes clear that it is impossible to estimate probabilistic causation without having previously introduced it via \( \eta \).

Resorting to Bayesian estimation does not change the problem. In the Bayesian approach the picture is essentially the same even if it is not so clear as we have just seen. First, we must return to random variables related to attributes and, secondly, in order to estimate probabilities we must introduce a prior distribution. Probabilistic causality is related to the choice of the prior.
3. What we have seen in the previous section suggests that we cannot estimate probabilistic causality. Obviously, deeper analysis is needed in order to exclude all possibility of performing such an estimation. In what follows, so as to throw some light on the matter, I consider two examples in which the existence of probabilistic causality can be checked. The first, related to statistical mechanics involves only one modality, i.e. one family of predicates. In this case the problem we are facing can be completely solved.

If in addition to the validity of regularity, symmetry and invariance, we suppose that both

$$p_j = 1/k$$

and

$$(6) \quad \eta = k/(k + 1)$$

hold, then we can prove that

$$(7) \quad \Pr\{X_i = j \mid E^n\} = \frac{n_j + 1}{n + k}$$

i.e. that Gini’s rule of succession holds. Using (7), via the product rule, we obtain the Bose Einstein statistics, i.e.

$$\Pr\{N\} = \left( \frac{N + k - 1}{N} \right)^{-1}$$

where \(N = (N_1, \ldots, N_k)\) is the structure description of a population of \(N\) bosons belonging to a cell endowed with \(k\) oscillators. If \(N\) and \(k\) go to infinity, so that the average number of bosons per oscillator \(N/k\) tends to \(\langle n \rangle\), then the distribution on states of an oscillator, \(m = 0, 1, 2, \ldots\), is

$$(8) \quad \Pr\{m\} = \frac{\langle n \rangle^m}{(1 + \langle n \rangle)^{m+1}}.$$  

in which the sole parameter is \(\langle n \rangle\).

Considering oscillators with different frequency \(\varepsilon_d\), and taking into account the entropy of the system, we can determine the value of \(\langle n_d \rangle\) for which the entropy of the system reaches its maximum. This maximum entropy distribution is that for which the average number of bosons (state) per oscillator is

$$(9) \quad \langle n_d \rangle = \frac{1}{\exp\left(\frac{\varepsilon_d - \mu}{kT}\right) - 1},$$

where \(\varepsilon_d\) is the frequency of the oscillator, \(\mu\) is the chemical potential, and \(k\) and \(T\) are the Boltzmann constant and temperature, respectively.
where $K$ is Boltzmann's constant, $T$ is the temperature and $\mu$ the chemical potential (Costantini and Garibaldi, unpublished).

In order to check (8) with $\langle n_d \rangle$ fixed by (9), we must sample the number of bosons per oscillator. In principle we should take different values of $m$ in independent trials and determine the (sampling) distribution of $m$. But in this picture the enormous velocity at which changes occur at the microscopic level is not taken into account. Due to the fact that the time resolution of the analyzer is much greater than the characteristic fluctuation time, an experimental value is actually a mean value, that is, a value of the practically deterministic function (9). Measurements at different frequencies give the blackbody radiation spectrum, i.e. a particular form of (9) with $\mu = 0$.

In this way we have checked that photons are positively correlated. In fact, what we have just performed is a significance test. Of course the null hypothesis is a composite one, i.e. the stated probability conditions, in particular (6). From these conditions we have derived the blackbody radiation. As a matter of fact, (9) conforms pretty well with the energy coming from a furnace. Hence, following a well-known statistical recipe, we can conclude that photons satisfy conditions used to derive (9). That is, besides other conditions, (6) also holds. In other words, the fact that a photon is on an oscillator is a prima facie quadrant cause of the presence of another photon on the same oscillator. In physical jargon, this is called 'stimulated emission' because the probability of accommodation of a photon on an oscillator in the state $m$ is proportional to $m + 1$.

What we have seen is essentially explained by the following. From the assumption of a probability relation between two (generic) individuals, i.e. the existence of a definite value of $\eta$, we have derived the frequency distribution of a population with respect to a modality. Having ascertained that the predicted distribution is actually the one we experimentally observe, we have asserted the existence of a probabilistic relation of causality among individuals of that population.

4. In the second example I consider two modalities. In this case difficulties arise from two sides. Firstly, we are not able to treat probability problems involving two modalities. As a consequence, we do not know exactly what we are doing when we are stating probability conditions referring to two modalities. Second, we must account for two types of causality. One, so to say horizontal, is of the same type we have
considered in the first example. That is, it involves random variables having predicates of the same family. The other, so to say vertical, is connected with the causality of one modality upon the other. Luckily, the example we are considering excludes both types of causality, but this is only due to the very coarse level at which the problem is usually faced by statisticians.

The problem amounts to checking the existence of a probabilistic causal relation with reference to two modalities. This is the case of the usual statistical practice. One of the most celebrated cases is that of the twin brothers and sisters of J. Lange studied by R. A. Fisher. (1958, Section 21.02) The evidence of Lange is that among 13 criminals who were monozygotic twins, 10 had twin brothers or sisters who had also been convicted. On the other hand, among 17 criminals who were dizygotic twins of like sex, two had convicted twin brothers or sisters. The statistical inference, named by Fisher as the exact treatment of $2 \times 2$ tables, amounts to checking whether these data show some probabilistic causal relationship or not. More precisely, whether data show criminality is more probable among monozygotic twins of criminals than among dizygotic twins of criminals.

Because we do not possess more precise reports about which brother or sister is or is not criminal, the data are that of the following table

<table>
<thead>
<tr>
<th></th>
<th>convicted</th>
<th>not convicted</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>monozygotic</td>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>dizygotic</td>
<td>2</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>total</td>
<td>12</td>
<td>18</td>
<td>30</td>
</tr>
</tbody>
</table>

Contrary to the case of photons, a precise value of $\eta$ cannot be suggested. As is usual in statistics, it is supposed that there is no prima facie quadrant cause: independence is assumed. But, due to the fact that I do not follow Fisher’s exact treatment, I must say more about this assumption.

First of all, we need two modalities: one with attributes ‘monozygotic’ and ‘dizygotic’, $F_1 = \{M, -M\}$, and another with attributes ‘convicted’ and ‘not convicted’, $F_2 = \{C, -C\}$. We also need 30 random variables, $X_1, \ldots, X_{30}$, one for each individual in the sample. We are interested in compound random processes, that is, in state descriptions going down and up from $F_1$ to $F_2$ and vice versa. The probability
of one of these state descriptions is

$$
\begin{align*}
\Pr\{X_1 = M\} \Pr\{X_1 = C \mid X_1 = M\}, \\
\Pr\{X_{30} = -M \mid X_1 = M, X_1 = C\}, \ldots ,
\end{align*}
\begin{align*}
X_{29} = -M, X_{29} = -C \\
\Pr\{X_{30} = -C \mid X_1 = M, X_1 = C, \ldots ,
\end{align*}
\begin{align*}
X_{29} = -M, X_{29} = C, X_{30} = -M\}.
\end{align*}
$$

Echangeability enables us to express this probability as follows:

$$
\begin{align*}
\Pr\{M \mid 0, 0 : 0, 0\} \Pr\{C \mid 1, 0 : 0, 0\} \\
\Pr\{M \mid 1, 0 : 1, 0\} \Pr\{C \mid 2, 0 : 1, 0\}, \ldots \\
\Pr\{M \mid 9, 0 : 9, 0\} \Pr\{C \mid 10, 0 : 9, 0\} \\
\Pr\{M \mid 10, 0 : 10, 0\} \Pr\{-C \mid 11, 0 : 10, 0\} \\
\Pr\{M \mid 11, 0 : 10, 1\} \Pr\{-C \mid 12, 0 : 10, 1\} \\
\Pr\{M \mid 12, 0 : 10, 2\} \Pr\{-C \mid 13, 0 : 10, 2\} \\
\Pr\{-M \mid 13, 0 : 10, 3\} \Pr\{C \mid 13, 1 : 10, 3\}, \ldots \\
\Pr\{-M \mid 13, 16 : 12, 17\} \Pr\{-C \mid 13, 17 : 12, 17\}.
\end{align*}
$$

The conditions we have stated say nothing about probabilities involving attributes of different modalities. I have no intention of facing the problem of determining probabilities related to two modalities because, in order to check independence, we only need the validity of the equality

$$(10) \quad \Pr\{\pm C \mid \pm M, E\} = \Pr\{\pm C\}$$

where $E$ is any piece of evidence of whatever nature. This condition is very far reaching. It excludes any type of dependence. The probability of $(\pm)C$ does not change knowing: the number of $(\pm)C$ in $E$; the number of $(\pm)M$ in $E$; the number of $(\pm)M$ which are $(\pm)C$; and, finally, that the considered individual is $(\pm)M$.

For the sake of simplicity we suppose that

$$(11) \quad \Pr\{X_i = M\} = \Pr\{X_i = C\} = 1/2, \quad 1 \leq i \leq 30.$$

This together with (10) gives the probability of the state descriptions we are considering. This probability is

$$(1/2)^{13}(1/2)^{17}(1/2)^{12}(1/2)^{18}.$$
Thus the probability of the table is

\[
(12) \quad \binom{13}{3} \binom{17}{15} (1/2)^{60}.
\]

The probability of one of the most probable tables, determined with the same assumptions we have just made, is

\[
(13) \quad \binom{15}{7} \binom{15}{7} (1/2)^{60}.
\]

The ratio between (12) and (13) is equal to 0.00095. We must be cautious in using this figure. The ratio points out that the observation is highly exceptional. Following Pearson, W. S. Gossett and Fisher, we must reject the conjunction of conditions we have used in order to derive (12). I am aware of the fact that, in this respect, it is possible to say much more than that which I am willing to say. This follows from the variety of suppositions stated in (10). But I have no intention of discussing this matter. At the same coarse level at which statisticians move, I direct my attention to modalities only. I affirm that there is a prima facie quadrant cause between the fact that one of our random variables has value \( M \) and that the same variable has value \( C \). Or, in other words, that the attribute \( M \) is a prima facie quadrant cause for the attribute \( C \). Briefly, that to have a convicted monozygotic twin causes – probabilistically – the conviction.

5. What I have done gives rise to a problem. The definition of quadrant cause is posed in terms of two random variables \( X_t \) and \( Y_t \). In Suppes’s original approach these variables refer to attributes. I have shown that quadrant causality may be treated by referring to the Gini–Johnson–Carnap tradition, too. That is, I have shown that it is possible to approach quadrant causality in the frame of predictive statistics. In this case the random variables refer to individuals. In the first example I have examined individuals bearing attributes of the same modality: in the second, attributes of two modalities. But the problem does not lie here. The problem I am thinking about arises from the fact that the hypotheses I have made – in one case causality, (6), in the other independence, (10) – refer to individuals, whereas the control has been made by taking frequency distributions into account. To avoid any misunderstanding I make clear that, due to the fact that in his definition
random variables refer to attributes, this is not a difficulty for Suppes's approach.

Equations (6) and (10) do not refer to definite individuals. I have assumed exchangeability, hence I have not considered well defined individuals. In other words, in both cases I have considered individual statistical laws. More exactly, (6) and (10) assert that the predictive probability that an unobserved individual bears an attribute, given that an observed individual bears another attribute, is greater than, respectively equal to, the predictive probability that an unobserved individual bears that attribute. The difficulty is due to the fact that, notwithstanding the fact that I am stating something about individuals, I must check this assertion by making use of frequency distributions.

This difficulty is not specific to problems dealing with probabilistic causality. The same problem arises in connection with any control of probability statements. Of course, for students sharing the frequency interpretation of probability, this is not a problem. Each probability statement is a frequency statement, and the control of such a statement can only be done by referring to frequency distributions. There is also no problem for students for whom the probability is a degree of belief. For a subjectivist no probability assertion can be checked with a frequency distribution owing to the very simple reason that, for him, there are no such distributions.

It is easy to realize that the problem I have in mind only arises for students sharing a logical interpretation of probability or, more generally, for objective Baysians. If one maintains that statistical inferences deal with individual probability statements which are not degrees of belief, then the problem arises. This is exactly what happens for the cases we have examined which are particular cases of statistical inferences based on individual statistical laws. Due to symmetry, no law of this sort can refer to a definite individual. In a sense, an individual statistical law asserts that a generic individual bears an attribute. However, each observed individual either bears or does not bear an attribute. Hence it is impossible to check individual statistical laws observing results of single experiments. The only way to control individual statistical laws is to try to falsify them by making use of a frequency distribution. The procedure of falsification of statistical laws is as old as Arbuthnot; Laplace makes a great use of it; and finally it becomes fully scientific with Pearson, Gosset and Fisher. The procedure is as follows: from the individual statistical law whose validity has been assumed, in statisti-
cal jargon the null hypothesis, we must deduce a universal statistical law – statisticians speak of the sampling distribution. Such a law is a distribution allotting probability values to each possible composition of a sample of a given size. It is this distribution that makes the control feasible. In the case in which assumptions speak about the microscopic level, there is no problem. As we have seen for the black body radiation, we check macroscopic phenomena and they do not differ from deterministic ones. This is what happen in statistical mechanics. In statistics we must use some criterion of falsification: in statistical jargon, of rejection. It is really amazing that for years philosophers have discussed the problem of the falsification of scientific laws in the case in which this falsification is trivial, and have stopped analysing the problem in the case in which it becomes interesting, that is to say, in the case of statistical laws. Obviously I am referring to the debate about Popper’s philosophy. As a matter of fact, the interesting case has already been solved before philosophers even begin to discuss the trivial one. Tests of significance are the solution.

In conclusion I stress the main steps of this solution. First of all, a statistical law can only be controlled by referring to a frequency distribution. Secondly, in statistical mechanics the control is similar to that of a deterministic law. Thirdly, if the members of the sample are macroscopic objects, we need a rule of falsification. The search for such a rule has been solved in practice, at least for as long as Pearson used tail-probabilities. Tails of this sort are tabulated in each statistical textbook. But after the profound and destructive criticism of H. Jeffreys, a theoretical solution to the problem of falsifying statistical laws is needed. It is high time to face this problem.

ACKNOWLEDGEMENTS

The author is deeply indebted to U. Garibaldi and S. L. Zabell for helpful discussions and valuable suggestions.

*Istituto di Statistica,*
*Università degli Studi di Genova,*
*C.so Paganini 3,*
*16125 Genova,*
*Italy*
REFERENCES


COMMENTS BY PATRICK SUPPES

There are very few papers on the statistical aspects of the testing of probabilistic causality. It is good to have for this volume such a paper by Domenico Costantini. Certainly in my own monograph on this subject I did not say much about the statistical aspects of making an inference to the existence or non-existence of a probabilistic cause. I am also positive about Costantini using the quantitative notion of quadrant causality I defined, and thereby dealing with random variables rather than events. As I have pointed out in the past on several occasions, philosophers cut themselves off to a considerable extent from the standard statistical literature on many topics by dealing with probability spaces and events rather than random variables.

Costantini develops the theory from the standpoint of a logical theory of probability, or what he also terms an ‘objective Bayesian’ viewpoint. In the spirit of this approach, and following the tradition that he identifies as that of Gini, Johnson and Carnap, he moves from attributes to individuals in the discussion of quadrant causality, and thereby brings the analysis down to a still more concrete level, closer to what may well be required in a given statistical analysis where individuals constitute the sample from which a sample distribution is constructed.

In spite of my agreement with many of the details of the analysis, I am not sympathetic with the logical tradition in probability or objective Bayesianism, and there are difficulties with this approach, as he brings out in his article. I will not try to enter into a technical discussion here but just reiterate that my problem with objective Bayesian views is that I can often think of some other information or belief that I want to take account of in the analysis. I suppose that given the choice I would prefer for statistical purposes a frequentist approach to the objective Bayesian approach, although my real preference is for a subjective Bayesian view.
Of course it does seem to me that in the first of the two extended examples Costantini gives, the one about bosons, he is really using a standard theoretical concept of probability as it arises in the relevant physical theory. The probability results reached in his equations (8) and (9) do not depend on, and are not the least bit sensitive to, any particular view of probability. What they depend upon are simply the formal properties of probability used in the standard formulation of the relevant statistical mechanical theory.

In the second example concerning tendency to criminality among monozygotic twins when one twin has already been convicted, I would have to say again that my own preferred analysis would be Bayesian with a subjective prior, but I think that because of the strength of the evidence, Costantini and I would reach very similar general conclusions.

Two final comments about the remarks at the end of the paper. I certainly share his view that philosophers should pay more attention to the falsification of statistical laws, and also his view that there are already a great many specific inference procedures for doing so. On the other hand I did not really understand his enigmatic remark about H. Jeffreys in the very last paragraph of the paper. It does not seem to me that there is a profound theoretical problem about falsifying statistical laws, but perhaps I am misunderstanding what he means to imply in this final remark.