

Causality and Causal Models

Causalità e Modelli Causali

Benito V. Frosini

Università Cattolica del Sacro Cuore di Milano

Riassunto: Questo articolo cerca di offrire una panoramica sul tema delle relazioni causali a partire dagli sviluppi intervenuti in ambito filosofico, poi estesi e formalizzati nel campo dell'analisi statistica multivariata. Nella rassegna filosofica vengono analizzati vari concetti di causa, e i tentativi di individuare le cause spurie. La parte centrale del lavoro si occupa dei grafi aciclici direzionali, che costituiscono lo strumento elettivo di molte indagini causali, e viene evidenziata la differenziazione fra condizionamento e intervento su una variabile. Particolare attenzione è dedicata alle procedure di randomizzazione e pseudo-randomizzazione, volte ad evitare possibili confondimenti. L'articolo si chiude invitando all'uso di notevole cautela nella modellazione dell'indipendenza condizionale e nella sua verifica empirica.

Keywords: Causality, Causal models, Directed acyclic graph, Confounder, Counterfactual.

1. A Short Historical Overview Until 1948

Human beings, and - beyond them - practically all manifestations of life on earth, animal and vegetable, have some knowledge and remembrance of some causal-effect relations which affect their vital neighbourhood; this is essential for their existence and flourishing: human beings know they are unable to breathe and survive under water, while a different learning has been memorized by fishes. Since the early utterances of pseudo-scientific assertions and law-like relationships, mankind has always been interested in recording what can be expected from past observations. A famous example is provided by the Sumerian civilization (3,000-2,000 B.C.); some written works of theirs are collections of omens, whose typical form is "*if x happens, then y will happen*": from one or more occurrences or signs, some future events are predicted. Of course, the scientific basis for these predictions was rather poor, if not absent; most omens were essentially based on analogies, and the *signs* were often interpreted as messages sent by the gods (see e.g. Götzte, 1947).

In the Aristotelian philosophical building, the doctrine of causes is central, although not particularly and analitically developed, if contrasted with the whole corpus of Aristotle's thought. As for other elliptical concepts, the meaning of the word "cause" is context-dependent in Aristotle's works (as well as in subsequent authors): *cause* and *effect* can be viewed as physical objects, events, properties of things or systems. Of the four kinds of causes postulated by Aristotle we shall retain only the following two: (a) the efficient cause, and (b) the final cause (Physica II 194b, 15-30).

Although Aristotle distinguishes between *cause* and *explanation*, the two concepts are strictly tied with an extension of the (deductive) syllogism, which many years later (more than two thousand years later) was turned to represent the paradigm of scientific explanation. He writes (about this "syllogism that yields scientific knowledge"): "The

premises must be the causes of the conclusion, better known than it, and prior to it; its causes, since we possess scientific knowledge of a thing only when we know its cause; prior, in order to be causes; antecedently known, this antecedent knowledge being not our mere understanding of the meaning, but knowledge of the fact as well” (*Analitica Posteriora* I.2.71b 29-33).

About the events that are neither necessitated nor happen as a rule, Aristotle maintains some possibilities of scientific investigation: “But chance also and spontaneity are reckoned among causes: many things are said both to be and to come to be as a result of chance and spontaneity” (*Physica* II, 195b, 31-33).

The influence of the Aristotelian thought has spread over many centuries, and its rigorous frame has greatly helped the advent of the Scientific Revolution around the XVII century, by the outstanding discoveries of Bacon, Descartes, Galilei, Kepler, Newton, and many others. Although the Aristotelian cosmological model was completely abandoned, his methodological approach remained an important lesson to be followed by almost all scientists. The quite rigid, mostly deterministic, meaning of cause according to Aristotle was destined to permeate the scientific achievements of the XIX century. Even the probability calculus was viewed just as a surrogate, or an approximation, of deterministic laws not yet discovered; the use of probabilities was deemed as an avowal of one’s ignorance.

Laplace (1840, pp. 2-3) expresses this viewpoint very clearly: “Tous les événements, ceux même qui, par leur petitesse, semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l’ignorance des liens qui les unissent au système entier de l’univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu’ils arrivaient et se succédaient avec régularité, ou sans ordre apparent; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie, qui ne voit en elles que l’expression de l’ignorance où nous sommes des véritables causes”.

Quite another viewpoint about the cause-effect argument was sustained, in the first half of the XVIII century, by David Hume (*A Treatise of Human Nature*, 1739). Hume took quite a radical stand concerning the cause-effect relationship: basically he states that the several occasions of everyday life, as well as the observations carried out for scientific purposes, in which we speak of a condition *A* as a cause and a condition *B* as its effect, bear no justification on the facts *in ipsa re*, but are simply based on our habit of observing *B* after having observed *A*. Anyway, the purpose of Hume is not to dismiss the quite useful status of the cause-effect relationship, but only to downgrade such status to a more conventional basis. According to Hume, this relationship is characterized by three essential conditions: (1) contiguity in space and time; (2) succession in time (a cause precedes its effect); (3) constant conjunction.

Much more radical was the purpose of Bertrand Russell (1913), who denied any scientific status to the cause-effect relationship; in Russell’s words, “There is no question of repetitions of the “same” cause producing the “same” effect; it is not in any sameness of causes and effects that the constancy of scientific law consists, but in sameness of relations”. Quite similar views were expressed by Karl Pearson (1900) in his famous and influential “The Grammar of Science”. Writes Pearson (1900, p. 8): “Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated. Science for the past is a description, for the future a belief; it is not, and has never been, an explanation”. Although an empirical scientist - it is believed - can hardly

subscribe these words, it must be admitted that much of the argument is a *nominal* argument, depending on how much we are willing to stretch or to loosen the definitions of concepts. Actually, one cannot deny that scientific explanation is not limited to the discovery of causal relations, and that sometimes a scientist is content with explaining *how*, and not *why*; however, the whole history of science is permeated by the research of *causes* of given *effects*, both for an abstract desire of knowledge, and for the gaining of mastery of physical forces and biological phenomena.

Russell' and Pearson's views have been without consequence, both for philosophers and - mostly - for scientists, with the unpleasant exception of statisticians. As Salmon (1984, p. 138) writes, "... to many scientists uncorrupted by philosophical training, it is evident that causality plays a central role in scientific explanation".

Although hasty and very concise, such a survey of pre-1948 philosophical thought about causality cannot leave out the contribution of John Stuart Mill (1843); his book "A System of Logic" has been one of the most influential works of the nineteenth century, and its bearing has spread as far as our days. I assume that the reader is acquainted with at least some knowledge of Mill's methods: (1) the method of agreement, (2) the method of difference, (3) the joint method of agreement and difference, (4) the method of residues, (5) the method of concomitant variation (see. e.g. Copi-Cohen, 1998, Chap. 14). Although nobody is still willing to subscribe Mill's claims that his methods are the basic tools for *discovery* (of causal relations) and canons of *proof*, nonetheless they are valued as very powerful, and constantly applied. The core of the question - which could never find a satisfying and general solution - is that "the methods can yield reliable results only when the hypothesis that has been formulated does correctly identify the circumstances that are causally relevant"; but "circumstances in the real world do not come wearing tags marked "relevant" or "irrelevant"." (Copi-Cohen, 1998, pp. 529, 528).

2. From C.G. Hempel to W.C. Salmon

2.1. Deductive-Nomological and Inductive-Statistical Explanations

On the turn of the century from the publication of Mill's System of Logic, no new interest about causality in the philosophical domain arose, if one keeps aside such snobbish and nominal claims as those by K. Pearson and B. Russell, quoted above. It was just after the issue of Hempel-Oppenheim paper in 1948 that the related questions of causal-effect relationship and scientific explanation reached new vigour. Salmon (1984, p. 21) writes: "one might almost say that 1948 marks the division between the prehistory and the history of the philosophical study of scientific explanation", although acknowledging the legacy by Aristotle and previous writers such as Arnaud's Port-Royal Logic and Laplace.

Hempel contribution is best appreciated by putting together the 1948 paper and the 1962 essay (see also the books of 1965 and 1966). In great synthesis, Hempel's proposals for scientific explanations are of the following two kinds, according to the explanation it is strictly deterministic or statistical. In his approach to *Deductive-Nomological* (D-N) explanations, Hempel calls *explanandum* the fact or the event to be explained, and *explanans* the set of other facts or conditions that provide the required explanation. This last set is subdivided into two subsets: (a) the first subset is composed of particular explanatory conditions (or contingent circumstances) $C_1, C_2, \dots, C_k$; (b) the second subset comprises one or more general (deterministic or universal) laws $L_1, L_2, \dots, L_n$. If the explanandum E is implied by the conjunction of all the propositions in the explanans, the

Deductive-Nomological explanation of E takes the form of a deductive argument like the following:

$$\frac{C_1, C_2, \dots, C_k}{L_1, L_2, \dots, L_n} \quad (1)$$

$$E$$

The explanans and the explanandum together constitute the explanation (see also Salmon, 1984, pp. 15-30).

If, among the general laws required to explain the fact E , at least one is statistical, Hempel maintains the same reference scheme (1), just substituting the *single* line separating the premises from the conclusion (signifying that the argument is deductive), with a double line - such as in (2) - and an associated probability p , meaning that the argument is only inductively correct, and the number p “expresses the degree of inductive probability with which the premises support the conclusion” (Salmon, 1984, p. 30). It is presumed by Hempel that p is *fairly close to one*. Hempel calls this new scheme *Inductive-Statistical* (I-S) explanation.

In the simple case of a random experiment R (accomplished by man or nature), yielding an event E with a high probability p , the statistical law could be summarized by the equality $P_R(E) = p$, so that the inductive argument could be written as follows:

$$\frac{\text{Experiment } R \text{ is carried out}}{P_R(E) = p} \quad (2)$$

$$E$$

It is worth-while to specify that the *inductive* probability associated with the double line could be somewhat different - according to Hempel - from the probability (presumably of frequentist kind) written in the second premise of the argument; of course, such a change would need justification.

That p - in the above argument - be close to one, appears a natural requirement in order to cautiously generalize the D-N explanation in the presence of statistical laws: under the premises, event E would no longer be certain, however we can expect its realization with great confidence. The acceptance, or not, of this conclusion needs some clarifications.

Turning to classical (deterministic) logic, it is known - and easily controlled - that if, in a material implication $a \Rightarrow b$ (a and b statements that can be true or false) we make the conjunction of a with another statement c , the truth value of the implication remains the same: $a \Rightarrow b \mid (a \wedge c) \Rightarrow b$. When the implication $a \Rightarrow b$ is not always true, we can define its probability as

$$P(a \Rightarrow b) = P(B \mid A) = P(A \wedge B) \cdot P(A) \quad (3)$$

being A the set of elements - belonging to a given population - for which the statement a is true (same definition for B and other sets). Now, if A is substituted by the conjunction $a \wedge c$, the probability of $(a \wedge c) \Rightarrow b$ becomes

$$P((a \wedge c) \Rightarrow b) = P(B \mid A \wedge C) = P(A \wedge C \wedge B) \cdot P(A \wedge C) \quad (4)$$

which can assume quite distant values from (3); anyway, no general relationship exists between the two probabilities (cf. Frosini, 2002, pp. 170-172).

Such essential indetermination of an event probability, which is strictly dependent on the conditioning circumstances, draws a separation line between deductive logic and inductive logic. In inductive logic the setting of a reference population, and/or suitable conditioning, becomes essential for getting meaningful probabilities.

These considerations suggest that a useful and informative configuration of the sample space be based on the exploitation of all *relevant conditions* (for the facts to be explained). In this line of thought, Hempel (1965, pp. 394-403) posits as a preliminary condition to the application of an I-S scheme the *requirement of maximal specificity*; such a requirement appears strictly analogous to the *requirement of total evidence* of Carnap (1962, pp. 211-213): the explanans must include and exploit all the evidence which is relevant for the explanandum. However, what can happen is that - at best - all the *known* relevant factors are included in the explanation scheme, but the cause of the event E could go back to an *unknown* factor; moreover, the requirement of maximal specificity - as actually applied - cannot assure that *only* relevant factors will be included. The central question remains that of *relevance*, already commented on about Mill's methods.

But it is especially the requirement of a high probability in the scheme (2) that is untenable. First of all, let us consider the classical counterexample given by the recovery from a cold (Salmon, 1984, p. 30): if people who have colds take a daily dose of vitamin C for a fortnight, a proportion near 100 per cent will recover; but such a factor is wholly irrelevant, because almost all colds clear up within two weeks regardless. Two further counterexamples seem decisive: (a) if the sample space is only constituted by the event E and its complement $\bar{E}$, with probabilities $P(E) = 0.99$ and $P(\bar{E}) = 0.01$, what can explain E - with high probability - explains all the same $\bar{E}$ - with low probability; (b) given the condition C , the effect E arises only in a small fraction of cases, e.g. 2 per cent; but it is known, also, that E can occur *only* in the presence of C (C is necessary for E , with $P(E | C) = 0.02$, while E is sufficient for C , with $P(C | E) = 1$); while it is true that - given the cause C - the effect E seldom comes out, it is an outstanding fact to recognize that C is the only cause of E (see Scriven, 1959; Jeffrey, 1969).

These and other minor criticisms have led Hempel to acknowledge that most of them were well founded, and to radically revise his I-S scheme, thus coming closer to the approach of Reichenbach-Salmon (cf. Salmon, 1984, pp. 89-90).

3.2. Statistical Relevance Explanation

The Statistical-Relevance (S-R) model of scientific explanation, proposed by Salmon (1965, 1984), following Reichenbach (1956), is in its essence quite simple: to establish a causal relation between an effect E and a putative cause C , we must make a comparison with (at least) two probabilities, one related to the presence of C , and the other related to its absence. Formally, a factor C is *statistically relevant* to the occurrence of E under circumstances A if and only if

$$P_A(E | C) \dots P_A(E) \tag{5}$$

or, which is the same,

$$P_A(E | C) \dots P_A(E | \bar{C}) \tag{6}$$

provided $P_A(C) > 0$. In other words, under circumstances A , E and C are not independent.

In applications it is important to distinguish a *positively relevant* factor C when

$$P_A(E | C) > P_A(E), \text{ or } P_A(E | C) > P_A(E | \bar{C}) \quad (7)$$

and a *negatively relevant* factor when

$$P_A(E | C) < P_A(E), \text{ or } P_A(E | C) < P_A(E | \bar{C}). \quad (8)$$

Such distinction is appreciated also in the philosophical theory of causality, as it is usually posited that *causes raise the probabilities of their effects* (Eells, 1991, p. 1). However, this is only a very rough characterization of the probabilistic causal relationship.

The first reason of caution is that the existence of a relation like (7), by itself, is never able to establish a cause-effect connection, no matter how much the presence of *C* raises the probability of *E*. The simple reason is that the relation between *C* and *E* is *symmetric* (for the sake of simplicity, in the sequel we drop the indication of the background circumstances, or conditions, *A*):

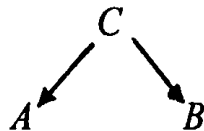
$$P(E | C) > P(E) \Leftrightarrow P(C | E) > P(C) \quad (9)$$

as easily checked; in contrast, a causal relation is clearly *asymmetric*. The usual and convenient way out for this impasse is simply - when available - the constant temporal succession of *C* and *E*, if we accept Hume's condition that a cause always precedes its effects. For the moment we can only observe that there could be a constant succession in time of *C* and *E*, thus displaying an apparent dependence, but that no causal dependence exists, owing to the influence on *C* and *E* of a "common cause", responsible for their dependence (an argument that immediately follows).

A second and more serious reason to stand *en garde* against the appearance of a causal relation, is the possibility that we have come across a *spurious causation* (or *spurious correlation*). Statisticians are well aware of what can sometimes happen in an analysis of regression: two variables *X* and *Y* display an unexpected connection, hardly ascribable to an underlying dependence; in most cases this is merely accidental, and we shall not dwell on this case any longer. In other cases, however, such a dependence is real, but it is not a causal relation: it is the case of *common cause* (or *conjunctive fork*).

The problem of "common cause" was thoroughly examined by Reichenbach (1956), and subsequently, constantly debated in the philosophical literature (see e.g. Salmon, 1984; Eells, 1991). A simple and recurrent example, taken from meteorology, is concerned with three successive events: (*C*) approaching cold fronts, (*B*) falling barometric reading, (*A*) raining. The events *A* and *B* are highly correlated, but no one would say that *B* causes, or explains, *A*; instead, they are both dependent - and caused - by the event *C*. In passing, we observe that a genuine cause relation is not necessary to make valid forecasts; the dependence relation between barometric reading and raining is quite sufficient for a reasonable (probabilistic) forecast, although there can be no causal dependence.

The reference graph for the common cause problem is as follows:



From Reichenbach (1956), the following conditions are assumed to hold for this model:

$$(1) P(A | C) > P(A | \bar{C})$$

- (2) $P(B | C) > P(B | \bar{C})$
 (3) $P(A | C, B) = P(A | C, \bar{B})$
 (4) $P(A | \bar{C}, B) = P(A | \bar{C}, \bar{B})$

Salmon (1978) calls this kind of probability structure a “conjunctive fork” (see also Eells, pp. 60, 107). The above conditions imply that $P(A | B) > P(A | \bar{B})$, so that an *apparent* causal relation could arise from the simple comparison of effects A and B , whereas the common cause C - if held fixed - screens off the correlated effects from one another. Formalizing the meteorological example, let A be the event “raining”, and B the event “falling barometric reading”, being the two events highly correlated; if C indicates “falling of atmospheric pressure” - independently of a particular and possibly mistaken barometric reading - the following equality ensues:

$$P(A | B \& C) = P(A | C) \quad (10)$$

i.e. C screens off B from A , or the events A and B become independent of each other, given C . A simpler writing - and a generalization - is achieved if we interpret A, B, C as Bernoulli (dichotomous) variables, by the use of the symbol $\$$ to mean independence; thus both conditions (3)-(4) of the conjunctive fork may be written $A\$B | C$.

Another famous example, widely commented on in the statistical literature, is Ronald Fisher (1959) proposal to consider the possibility that lung cancer (B) is positively correlated with smoking (A) not because of a direct causation from A to B , but owing to the existence of a genetic factor C (common cause), “able” to explain both phenomena (see e.g. Spirtes *et al.*, 2000, pp. 239-249).

A third, very important reason, not to rest on the first appearance of (real) dependence between two variables - or sets of variables, bears on cases collectively included under the name of *Simpson’s paradox* (Yule, 1903; Simpson, 1951; Eells, 1991, Chap. 2; Edwards, 2000, pp. 8-11; Spirtes *et al.*, 2000, pp. 38-42; Frosini, 2001, pp. 263, 290-292). The following simple example mimics the investigation made by Bickel, Hammel, and O’Connell (1977) on graduate admissions at Berkeley (numeric example taken from Eells, 1991, p. 63). Let us assume that the entries in the following table are the number of accepted applicants over the number applying to two departments.

	Department 1	Department 2	Total
Male	81/90	2/10	83/100
Female	9/10	18/90	27/100

If one looks at the overall figures (last column), the large proportions of male applicants accepted with respect to females is impressive, naturally suggesting discrimination against women. However, taking into account that admission decisions are made within the academic departments to which the person has applied for, it is checked that the proportions of male and female applicants who are accepted are the same within both departments, thus no discrimination has taken place. In a sense, “being male” is a positive cause for overall admission, but only because males preferably apply to a department that

is relatively easy to get into.

One could easily modify the above figures, showing that everything can occur, for example a discrimination against males in each department is compatible with no overall correlation between admission and sex. In conclusion, while correlation is blatantly no proof of causation, also causation does not imply correlation - overall, or on average.

All this is quite reasonable and well known; we could perhaps think of a case of N individuals, showing an *overall* positive (negative) correlation between variables X and Y , however with a negative (positive) correlation inside natural strata of the same individuals (see e.g. Frosini, 2001, p. 291). Besides, the problem of correctly identifying the classes of individuals for which a given causal - or dependence - relation makes sense, largely bears on Fisher's suggestion of conditioning on a convenient *reference set* (usually defined by the recourse to some ancillary statistic, in classical statistical inference - cf. Frosini, 1999).

2.3. Spurious Causes According to P. Suppes

One of the earliest and clearest investigations about the formal definition of genuine and spurious causes is contained in the famous monograph by Patrick Suppes "A Probabilistic Theory of Causality", of 1970. For typographic convenience, by $r < s < t$ we mean three successive times (instead of $t'' < t' < t$, used in the original work). Suppes starts by defining B_s a *prima facie* cause of the event A_t if $P(B_s) > 0$, and $P(A_t | B_s) > P(A_t)$, thus agreeing with the traditional approach expressed by the inequalities (7), and the idea that causes precede their effects.

Suppes writes (1970, p. 21): "The intuitive idea of spurious cause is that *an earlier event* may be found which accounts for the conditional probability of the effect just as well" (italics added). Two definitions of spurious cause are preceded by a *preliminary definition*:

Let B_s be a *prima facie* cause of A_t . Then B_s is a spurious cause of A_t if and only if there is a $r < s$ and an event C_r such that $P(B_s C_r) > 0$ and

$$P(A_t | B_s C_r) = P(A_t | C_r). \quad (11)$$

This equality is the same as (10), with the additional requirement of temporal precedence. A *prima facie* cause that is not spurious is called a *genuine cause*.

Suppes strongly maintains the temporal inequalities $r < s < t$, because otherwise - in his opinion - every cause could be declared spurious. The argument goes on as follows. Letting $r = s$, and $C_r = C_s$, if we put $C_s = B_s$ (B_s *prima facie* cause of A_t , with $P(B_s) > 0$),

$$P(A_t | B_s B_s) = P(A_t | B_s). \quad (12)$$

By comparing (11) and (12), Suppes thinks he has proved that - if $r = s$ and $C_s = B_s$ - "every cause is spurious". Perhaps that could formally hold, but this can blatantly appear at variance with the sentence quoted above, if we only replace *earlier* by *other*:"The intuitive idea of spurious cause is that *another* event may be found ...". Actually, what Suppes has done is that the *prima facie* cause B_s is screened off by itself, which seems untenable.

Although acknowledging that the above approach is somewhat forced, Suppes insists by giving another case, not asking for the unnatural equality $B_s = C_s$. Letting $C_s = B_s D_s$, being D_s distinct from B_s , such that $P(B_s D_s) > 0$, the following equality is obtained (noting that $B_s C_s = B_s D_s = C_s$):

$$P(A_t | B_s C_s) = P(A_t | C_s). \quad (13)$$

By comparing (13) and (11), Suppes states that “every event is spurious”. Also this interpretation is untenable: a subset C_s of the putative cause B_s screens off B_s , declaring it a spurious cause! In our opinion, no constrained time precedence can be assumed in the search aiming at characterizing prima facie causes as genuine causes.

After the “preliminary definition”, Suppes proposes two definitions of spurious cause; of these, only the latter appears of practical interest and well motivated. The idea is to identify a suitable partition of the actions - rather than the prima facie cause - which possibly brought on the effect A (for the above reasons, we do not yet write the time reference). Given a prima facie cause B , and a partition $\mathbf{B}$ such that for all elements C_k of $\mathbf{B}$ ($k = 1, 2, \dots$) $P(BC_k) > 0$, B is a spurious cause of A if and only if

$$P(A | BC_k) = P(A | C_k) \text{ for } k = 1, 2, \dots \quad (14)$$

As an example, let us consider the event A = lung cancer, for the workers of an industrial plant, and as a prima facie cause just the exposure of the workers to the particular kind of manufacturing process (event B), such that $P(A | B) > P(A)$; in other words, the proportion of cases of lung cancer for a large number of workers in the given industry is greater than the analogous proportion holding in the whole population (conditional on given sex and age distribution). Besides, let us assume that a research on the smoking habits of workers has revealed a quite larger consumption of cigarettes with respect to the population as a whole. If cigarette consumption is classified e.g. in four classes: C_1 = non smoking, C_2 = moderate smoking, C_3 = average smoker, C_4 = heavy smoker, and if the proportion of lung cancer in the four classes calculated on the workers (left hand side of (14)) and on the general population (right hand side of (14)) are coincident, then we can conclude that the prima facie cause B is spurious, and is screened off by smoking.

3. INUS Condition and Factorial Causation

To sum up, everything has been said so far, and most of what will follow, maintains a precise reference to *type-level causation*. i.e. to general kinds of events, phenomena, properties, factors; and actually, the typical background was *frequentist* probability. This is the most fruitful context for scientific investigations, as scientists aim at establishing *causal* relations between *types* of events. However, in life we are usually faced with problems of *token causation*, i.e. finding out “a relation between two actually occurring, concrete, token events” (Eells, 1991, p. 278). We shall not delve into the wide domain of token, or singular, causal claims, which has been the object of many philosophical studies (see e.g. Eells, pp. 278-392); instead, we can simply start from a proposal mostly suited for token causation, aiming at presenting an original theory of causation, recently issued in the statistical literature.

For an implication $A \Rightarrow B$ in deductive logic (A and B events described by sentences), it is usual to speak of A as *sufficient* for B , and B as *necessary* for A , and also of A and B as *necessary and sufficient* for each other in the case of double arrow $A \Leftrightarrow B$. But in many situations the causal relation is slightly more complicated: many (possible) sufficient causes are in the background, and most of them are not atomic, but derive from the conjunction of many atomic causes. Concerning causes of a specific disease, the

epidemiologist K.J. Rothman (1976; see also Rothman and Greenland, 1998, pp. 8-12) suggested a causal model establishing a biconditional statement between the disease D and the alternative between sets of (conjoint) sufficient causes. For example, if A, B, C, U are atomic factors, the following equivalence could be settled with the disease D :

$$D \Leftrightarrow ABU \text{ or } ACU \text{ or } BCU \quad (15)$$

where “or” is to mean an inclusive disjunction. In order not to include irrelevant factors, Rothman asks for including the *minimal* set of factors, in the sense that if we delete any of the factors making up the conjunction, what is left is not sufficient.

Along this same line of thought is J.L Mackie’s earlier proposal (1965, 1974), concerning the so-called *INUS condition*, meaning that the cause - or condition - C is an Insufficient but Necessary part of an Unnecessary but Sufficient cause of D if and only if there exist factors A and B such that

$$D \Leftrightarrow A \text{ or } (B \text{ and } C) \quad (16)$$

i.e. the condition A , as well as $(B \text{ and } C)$, are separately sufficient for D , while C is a necessary part of $(B \text{ and } C)$; lacking the conjunction with B , C is no more sufficient.

In order to implement probabilities, and treating $A, B, C, \dots$ as event sets, the following equalities and inequalities must be satisfied (Suppes, 1970, p. 76):

$$\begin{aligned} P(ACBC) &> 0 \\ P(D \mid ACBC) &= 1 \\ P(D \mid \overline{A \cup BC}) &= 0 \\ P(D \mid B) &< 1 \\ P(D \mid C) &< 1. \end{aligned}$$

Following Anderson (1938), Mackie assumes the existence of a *causal field* (S, F) , S being constituted by a set of objects or “opportunities” at which events might occur, and F being a set of factors, which separately or jointly can produce a given effect.

In view of enlightening the linkages with Rothman’s approach, if we are interested in focusing on C in equation (15) as a cause of D , such an equation may be rewritten as

$$D \Leftrightarrow ABU \text{ or } ((AU \text{ or } BU) \text{ and } C).$$

Among the problems of token causation, such an approach seems especially suited for civil and criminal cases in courts, where it is important to ascertain whether C is *contingently necessary*, in the sense that - in the given circumstances - the effect D would have not occurred had C failed to occur (Frosini, 2002, pp. 2-6).

Starting from the above ideas, Mikel Aickin (2002) has built a reference system of properties and structural equations able to manage - at least in principle - most problems in the domain of causality, and especially the problems most often encountered in biomedicine and epidemiology. The basic tools employed by Aickin are elementary, but sufficiently powerful: an equivalence like (16) between events (or propositions defining events) is translated by means of corresponding indicators; thus a random variable d takes

the value 1 if event D occurs, and the value 0 if it does not occur. The equations between indicators are not written using the ordinary algebra, but exploit the useful properties of so-called “unitary algebra”, which mimics corresponding relations in set theory. This algebra applies to *unitary variables*, taking values in the interval $[0,1]$, and *binary variables*, taking values 0 or 1. The basic properties of unitary algebra are defined as follows (Aickin, 2002, pp. 31-32):

$$\begin{aligned} u^* &= 1 - u \\ uv &= \text{product of } u \text{ and } v \\ uWv &= u + v - uv \\ u \setminus v &= (u - v) \cdot (1 - v) \quad \text{with } u \setminus 1 = 0. \end{aligned}$$

The operational meaning is immediately captured; for example, if u and v are indicators of events U and V , for example, $uv = 1$ if and only if $u = 1$ and $v = 1$ (both events occurring); $uWv = 0$ if and only if $u = 0$ and $v = 0$ (both events *not* occurring).

A first example of application is the parametrization of a simple probability model, i.e. the conditional expectation $E(y | x_1)$, being the bivariate variable of indicators y and x_1 defined on the field (S, F) :

$$E(y | x_1) = P(y = 1 | x_1) = \$_{0,1}W\$_1x_1 \quad (17)$$

where implicitly

$$\begin{aligned} P(y = 1 | x_1 = 0) &= \$_{0,1} \\ P(y = 1 | x_1 = 1) &= \$_{0,1}W\$_1 \end{aligned}$$

and then

$$\$_1 = P(y = 1 | x_1 = 1) \setminus P(y = 1 | x_1 = 0).$$

This same modelling process is easily extended to two or more “explanatory” variables; for example, (Aickin, 2002, p. 35):

$$P(y = 1 | x_1, x_2) = \$_{0,12}W\$_{1,2}x_1W\$_{2,1}x_2W\$_{12}x_1x_2 \quad (18)$$

(for the operational definition of parameters, the reader is referred to Aickin’s book).

While (17) and (18) can be called *fixed-effects models*, if $b_{0,1}$ and b_1 are binary variables we can write a *random-effects model* such as

$$y = b_{0,1}Wb_1x_1 \quad (19)$$

(and a similar equation in the case of two variables x_1 and x_2 , replacing $\$$ with b in (18)). Equation (19) clearly parallels equivalence (16): y is the effect indicator, and $y = 1$ occurs either if $b_{0,1} = 1$ (residual cause) or if $b_1x_1 = 1$, which takes place when $x_1 = 1$ is supported by the co-cause $b_1 = 1$ (and of course when all indicators equal 1). Only in special cases model (19) implies the simple conditional expectation (17): the simplest case requires the conditional independence $b_{0,1} \$ b_1 | x_1$ (cf. Aickin, 2002, p. 37).

As an example (from Aickin, 2002, p. 59), let us assume that two variables R and C , each taking three values, can be responsible for the effect y ; calling r_1, r_2, r_3 the indicators of the R events, and c_1, c_2, c_3 the indicators for C events, let $x_1 = r_3$ (the third row of the 3×3 contingency table is especially investigated as a cause of y). Further on, let $y = 1$ occur in the presence of the couples: $r_1c_1, r_2c_2, c_2x_1, c_3x_1$. The causal equation of type (19) may

thus be written:

$$y = (r_1 c_1 W r_2 c_2) W (c_2 W c_3) x_1$$

from which a useful diagram may be constructed (Aickin, 2002, p. 59).

Given a causal field $(\mathbf{S}, F)$, we are in a position to write a similar causal equation, provided the set F is sufficiently large, possibly including unobservable factors. Formally, the factors $b_{0,1}$ (residual cause) and b_1 (co-cause) can thus be determined by the following stages - and similarly in cases of two or more factors. First of all, we single out the pathways (conjunctions of factors $f \in F$) not containing x_1 which imply the occurrence of the given effect, or $y = 1$: the alternation of these pathways gives rise to $b_{0,1}$. Next we can examine the pathways, not considered in the previous alternation, and form the alternation of those pathways which imply the effect, when joined with x_1 : such an alternation gives rise to b_1 . Even if it is so loosely defined (especially when unobserved variables are involved), such a causal equation allows to extract important pieces of information from the causal field $(\mathbf{S}, F)$, if supplemented with a probability distribution over $\mathbf{S}$. In fact, we can always define parameters $\$_{0,1}$, $^*_{0,1}$ and $\$_1$ as follows:

$$P(b_{0,1} = 1 \mid x_1) = \$_{0,1} W^*_{0,1} x_1 \quad (20)$$

$$P(b_1 = 1 \mid x_1 = 1, b_{0,1} = 0) = \$_1 \quad (21)$$

and then derive the causal equation (Aickin, 2002, p.50):

$$P(y = 1 \mid x_1) = \$_{0,1} W(\$_1 W^*_{0,1} x_1) \quad (22)$$

From (20) and (21) one obtains

$$\begin{aligned} \$_{0,1} &= P(b_{0,1} = 1 \mid x_1 = 0) = P(y = 1 \mid x_1 = 0) \\ \$_{0,1} W^*_{0,1} W \$_1 &= P(y = 1 \mid x_1 = 1) \end{aligned}$$

thus from observations of random draws from $\mathbf{S}$ it is possible to estimate only two parameters out of three. One fortunate - or forced - condition is $b_{0,1} \$x_1$, namely the residual cause is independent of the factor x_1 ; in fact, it is immediately checked that such independence implies, and is implied by $^*_{0,1} = 0$. Thus the important result is

$$b_{0,1} \$x_1 \Leftrightarrow P(y = 1 \mid x_1) = \$_{0,1} W \$_1 x_1 \quad (23)$$

which holds under the IFR condition (Independence of the Factor and the Residual cause). A traditional way to ensure such independence - when possible - requires the implementation of a *randomized* mechanism, through which the presence or absence of the factor x_1 on the sampled units (or the population units) is determined by fixing a certain probability for the factor, irrespective of the units. For further developments of these ideas the reader is heartily referred to Aickin's book.

4. Directed Acyclic Graphs as Causal Models

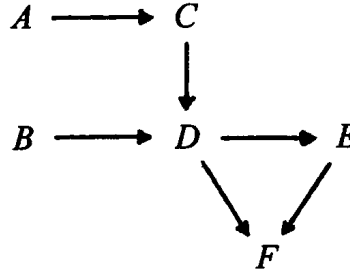
4.1. Directed Acyclic Graphs and Related Distributions

Among the several graphical models introduced to manage the relations between random

variables, *directed acyclic graphs* - DAGs for short - seem the ones most suitable to exploit and develop most of the potentialities concerning the concept of *cause*, as introduced in previous sections. Examples of such graphs were perhaps first introduced by Wright (1921, 1934), in his proposal of *path analysis*. Subsequent applications were devised for categorical data (L.A. Goodman, 1973; Wermuth and Lauritzen, 1983), but it has only been in the latest two decades - mostly from 1993 - that interest in causal models and DAGs has aroused an explosion of writings, articles and monographs.

A directed acyclic graph, or DAG, is something like Figure 1, representing *formal* causal relationships between six variables, named $A, B, \dots, F$. In the sequel, only discrete and finite variables will be contemplated, unless otherwise stated. Thus a DAG G is formally defined as a pair $G = (V, E)$, being V a finite set of *vertices*, or *nodes*, representing variables, and E the set of directed edges - or arrows; the set E is a subset of the cartesian product $V \times V$. A DAG is acyclic in the sense that no loops are allowed; e.g., if the direction of the arrow from D to F is reversed, the graph would cease to be a directed *acyclic* graph.

Figure 1: An example of directed acyclic graph with six variables



An arrow, or directed edge, between two variables, means the existence of some (causal) direct influence between the variables involved; for example, $A \rightarrow B$ means that the variable A has an influence on variable B , or - in other words - that $P(B = b_j | A = a_r) \dots P(B = b_j | A = a_s)$, for $r \dots s$, at least for some j, r, s .

A fundamental and useful terminology applied to DAGs borrows from family relationships: if $A \rightarrow B$ we say that A is a *parent* of B , or that B is a *child* of A . Even, if there is a *directed* path from A to B (possibly including other nodes), then A is called an *ancestor* of B , and B is called a *descendant* of A . The terminology taken from econometrics can also be useful: the variables with no inputs, i.e. with no parents, are called *exogenous*; all the other variables are called *endogenous*. Loosely speaking, the scheme in Figure 1 represents the influence displayed by two independent and exogenous variables A, B on the final effect F , mediated by three endogenous variables C, D, E .

The *skeleton* of a graph $G = (V, E)$ is its underlying undirected graph, simply obtained from G by letting the edges without arrows. Given three distinct vertices A, B, C of G , the triple (A, B, C) is called an *immorality* of G if the subgraph containing these nodes is of kind $A \rightarrow B \leftarrow C$, i.e. if the parents A and C of B are “unmarried”. An equivalent terminology calls B a *collider* if it has converging arrows from A and C .

Apart from the simple influence relationship between couples of variables connected by a directed edge, a DAG like Figure 1 is characterized by the factorization of the conjoint density, according to the presence and direction of edges, generally yielding several kinds of reduction with respect to the standard factorization of a generic density, such as

$$p(x_1, \dots, x_n) = \prod_j p(x_j | x_1, \dots, x_{j-1}).$$

For example, the density relating to the graph in Figure 1 may be written as follows:

$$p(a,b,c,d,e,f) = p(a)p(b)p(c | a)p(d | b,c)p(e | d)p(f | d,e)$$

where the factors on the right hand side - for the sake of simplicity - are the densities of the respective variables, possibly conditional on the values of their parents.

More generally, calling $pa(A)$ the collection of parents of A (which is void in the case of exogenous variables), with the definition

$$pa(A) = \bigwedge_{0_A} pa(\cdot) \setminus A$$

in the general case of a set A containing 0, 1, 2, ... nodes, the factorization of the joint density of the k nodes in the DAG G (corresponding to variables $X_1, \dots, X_n$) can be written as

$$p(x_1, \dots, x_n) = \prod_j p(x_j | pa(X_j)). \quad (24)$$

While a DAG G allows a characterization of the joint density such as (24), many, or even all the arrows (directed edges) included in a DAG, cannot be derived from the knowledge of the joint density; this will arise as a comment of the following elementary constituents of a DAG.

I) $A \rightarrow B$

may be interpreted as “ A is cause of B ”, with density factorizing as $p(a,b) = p(a)p(b | a)$; note however that the same density is reflected by reversing the arrow:

$$A \leftarrow B \quad \text{with density } p(a,b) = p(b)p(a | b) = p(a)p(b | a).$$

II) (a) $A \rightarrow B \rightarrow C$

A is a cause of C , mediated by B : $p(a,b,c) = p(a)p(b | a)p(c | b)$.

Just as before, it is easy to check that the following DAGs yield the same joint density:

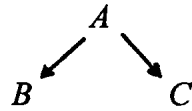
$$\text{II (b)} \quad A \leftarrow B \rightarrow C \quad ; \quad \text{II(c)} \quad A \leftarrow B \leftarrow C$$

For DAG II(a), the intermediate node B has A as parent and C as child. If we write the probability of (A,C) conditional on B , we obtain

$$p(a,c | b) = p(b)p(a | b)p(c | b) \cdot p(b) = p(a | b)p(c | b)$$

i.e. conditioning on the intermediate cause B , the first cause A and the effect C become conditionally independent.

III) (a) If the DAG reflecting the causal structure for variables A, B, C is II(b), it is usually displayed as (with different roles for A, B, C):

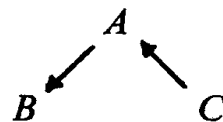


and this is the scheme of the *common cause*, extensively dealt with in Section 2.2. As $p(a,b,c) = p(a)p(b|a)p(c|a)$, it is immediately checked that the effects B and C become conditionally independent given the common cause A . Of course, from the point of view of the density, the following graphs are equivalent:

III(b)



III(c)



IV) The *common effect* of two independent causes may be represented as $A \rightarrow C \leftarrow B$ with joint density $p(a,b,c) = p(a)p(b)p(c|a,b)$.

In this case, reversing one or both the arrows, the graph is turned into one of the graphs of kind III above, which possess a factorization of the density not reducible to the one just written. An interesting feature of the common effect scheme is that the (marginally) independent causes A and B become dependent by conditioning on the effect. As a simple example, let X and Y be the outcomes of two dice, each with points $1, \dots, 6$, and $Z = X+Y$. Conditioning on $Z = 4$, the following bivariate distribution for $X, Y | Z = 4$ obtains:

$Y Z = 4$			
$X Z = 4$	1	2	3
1	0	0	1/3
2	0	1/3	0
3	1/3	0	0

Two important questions need to be answered, when considering much larger graphs than the ones previously examined. The first relates to the family of graphs which are *equivalent* from the probabilistic viewpoint, however with at least one reverse direction with respect to the other graphs of the family. The characterization of this family is given by the following theorem (Verma and Pearl, 1990; Frydenberg, 1980, Andersson *et al.*, 1997):

Two DAGs are (Markov) equivalent if and only if they have the same skeleton and the same immoralities.

A consequence of this theorem is that the number of equivalent DAGs grows very rapidly with the number of vertices, soon becoming intractable even with the fastest computers. As all these causal models are probabilistically equivalent, observation or estimation of the structure of the joint density can never lead to a knowledge of the causal structure. When feasible, a simple criterion to follow is to settle the direction of arrows according to time ordering, thus applying the most intuitive of Hume's criteria. Of course, in itself such a criterion is worthless: even if we accept the time succession between cause

and effect, recognition of a variable A as a (*direct*) cause of a variable B is hardly limited to recognition of time succession and bare exclusion of independence.

The second question entails the characterization of DAGs according to their independence structure, both marginal and conditional; this is dealt with in the next subsection.

4.2. Markov Properties of Directed Acyclic Graphs

One of the most important concepts used to model multivariate data is independence, marginal and conditional. Following Dawid (1979), we write $X \perp\!\!\!\perp Y \mid Z$ meaning that in the joint probability distribution of the random variables X, Y, Z the probability of $(X = x, Y = y)$, given $Z = z$, factorizes in

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

assuming that $P(Z = z) > 0$. Equivalent characterizations concerning densities are:

$$f(x, y, z) = f(x \mid z)f(y \mid z) \quad (25)$$

$$f(x, y, z) = h(x, z)k(y, z) \quad \text{for some } h, k; \quad (26)$$

other characterizations may be found in Dawid (1979) and Lauritzen (1996).

Some conditional independencies were already singled out in commenting on the simple graphs in Subsection 4.1; however, many other independencies are implied by a more complex graph and its associated density factorization of kind (24). The easier to recognize are perhaps the most interesting ones: each variable is independent of its non-descendant conditional on its parents (see e.g. Dawid, 2002, p. 162); this property is usually called *Causal Markov Condition* (Pearl, 2000, p. 30; Spirtes *et al.*, 2000, p. 29). Other independence properties in a complex DAG have been investigated by Verma and Pearl (1990) and Lauritzen *et al.* (1990).

4.3. Directed Acyclic Graphs and Structural Equation Models

The (formal) causal relationships expressed by directed edges in a DAG simply aroused from the general definition of “dependence” as “non independence”, thus comprising practically all conceivable relations between a variable and its parents. However, in many fields of applied research (e.g. econometrics, engineering, genetics, social sciences), such relationships are expressed in a *deterministic* way, by means of exact functional relationships, meaning that “for each assignment of a unique value to each of the parent vertices, there is a unique value of the dependent vertex” (Spirtes *et al.*, 2000, p. 15). Concerning e.g. the graph in Figure 1, and for suitable choice of functions f_C, f_D, f_E, f_F , to this graph could correspond the structural equation model

$$\begin{aligned} C &= f_C(A) \\ D &= f_D(B, C) \\ E &= f_E(D) \\ F &= f_F(D, E) \end{aligned}$$

In general, letting $X_1, \dots, X_n$ the variables (vertices), and $pa(X_i)$ their parents, a structural equation model is given by a set of equations of the form

$$X_i = f_i(pa(X_i)) \quad i = 1, \dots, n \quad (27)$$

(cf. Pearl, 2000, p. 27). In many applications, if $X_1, \dots, X_n$ are observable variables, such equations are usually defined by adjoining to their parents an *unobservable* or *error* variable U_i ; these variables are exogenous, hence mutually independent; the above model thus becomes:

$$X_i = f_i(pa(X_i), U_i) \quad i = 1, \dots, n. \quad (28)$$

This is precisely the type of structure usually assumed in econometric models, except that in some models the above structure contemplates the existence of feedback reactions (see Faliva, 1992; Faliva and Zoia, 1994).

4.4. The Intervention Calculus

As already stressed in Section 4.1, the connections between a directed acyclic graph - with arrows meaning some kind of causal relationship - and the associated probability distribution, are rather loose, as many different directions of edges are compatible with the same distribution. Lacking a stronger condition, able to establish a firm connection between the direction of edges in a DAG and its probabilistic modelling, any definition of a causal model would just be fancy. Such a condition exists, and has already been skimmed just in the preceding lines about structural equations.

Turning to the simple graphs $A \rightarrow B$ and $B \rightarrow A$, it has already been stressed out that they are indistinguishable from the point of view of the joint density, as $p(a, b) = p(a)p(b | a) = p(b)p(a | b)$. However, something different will happen if we *decide* (or some external force determines) that $A = a$ - one of the possible outcomes of the variable A . Note that this is quite different from the observation that $A = a$, i.e. produced by the random mechanism summarized by $p(a, b)$; thus we need some symbol expressing the fact that some variable is conditioned *by intervention* - or *by action* - on another variable, as contrasted to the symbol “|” used for conditioning *by observation*. Following Lauritzen (2001, p. 76), we shall make use of the symbol “ $\leftarrow$ ”, thus writing

$$p(b \downarrow a) = P(B = b | A \leftarrow a)$$

where $A \leftarrow a$ expresses the condition that the variable A is *forced* to take the value a (the analogous symbolism introduced by Pearl (2000, p. 70) is “ $do(A = a)$ ”).

Making use of this intervention symbology, if we force A to take the value a we may write, with reference to the graph $A \rightarrow B$:

$$\begin{aligned} p(b \downarrow a) &= P(B = b | A \leftarrow a) = P(B = b | A = a) = p(b | a) \\ p(a \downarrow b) &= P(A = a | B \leftarrow b) = P(A = a) = p(a) \end{aligned}$$

and with reference to the graph $B \rightarrow A$:

$$\begin{aligned} p(b \downarrow a) &= P(B = b | A \leftarrow a) = P(B = b) = p(b) \\ p(a \downarrow b) &= P(A = a | B \leftarrow b) = P(A = a | B = b) = p(a | b) \end{aligned}$$

thus showing an asymmetry between intervention and (simple) conditioning: the intervention calculus is influenced by the direction of the arrows in a DAG, while

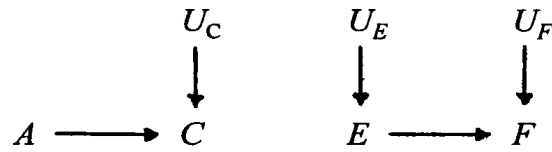
conditioning is not.

The above considerations are especially easy to apply when a DAG has a correspondingly structural equation model (or recursive equation system); in the case of the system examined above in this section, inserting - as usual - the error terms as in (28), if D has been set to d the model is reduced to:

$$\begin{aligned} C &= f_C(A, U_C) \\ E &= f_E(d, U_E) \\ F &= f_F(d, E, U_F) \end{aligned}$$

with a corresponding graph (cf. Figure 1):

Figure 2: *A modification of Figure 1 by intervention*



If we exclude an explicit introduction of the unobserved parents U_i for each node, setting $D = d$ changes the equation system as follows: $C = f_C(A)$, with the graph reduced to $A \rightarrow C$; all the other nodes have been set equal to constants, according to their relations (direct and indirect) with $D = d$.

Before going on, a warning is necessary. In introducing and explaining the effects of interventions in the simple situations above, we have maintained that the overall probabilistic mechanism is not altered by the intervention, which restricts its action to cutting all links of the nodes involved from their parents, and determining their descendants as the *given* functions of the set values - and possibly of other variables. This is by no means obvious, or taken for granted; on the contrary, it is a precise assumption that must be explicit, and - in any application - needs to be justified; “subject matter knowledge must play an essential role in this justification process” (Lauritzen, 2001, p. 95).

Having said that, if the factorization (24) holds (i.e. if the distribution P is Markov with respect to the associated DAG G), the directed acyclic graph G is said *causal* if, under an atomic intervention $X_i \leftarrow x_i^*$,

$$p(x_1, \dots, x_n \mid x_i^*) = \begin{cases} \prod_{j \dots i} p(x_j \mid pa(X_j)) & \text{if } x_i = x_i^* \\ 0 & \text{if } x_i \neq x_i^* \end{cases} \quad (29)$$

Multiplying and dividing (29) by $p(x_i^* \mid pa(X_i))$, the following equivalent formula obtains:

$$p(x_1, \dots, x_n \mid x_i^*) = p(x_1, \dots, x_n) \cdot p(x_i^* \mid pa(X_i)) \text{ with } x_i = x_i^*$$

(cf. Pearl, 2000, pp. 72-73; Lauritzen, 2001, p. 76), and also, if we interpret the division by $p(x_i^* \mid pa(X_i))$ as conditionalization on x_i^* and $pa(X_i)$:

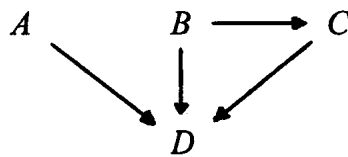
$$p(x_1, \dots, x_n \mid x_i^*) = p(x_1, \dots, x_n \mid x_i^*, pa(X_i)) p(pa(X_i)) \text{ with } x_i = x_i^*.$$

If we want to compute the effect of an intervention of kind $X_i \leftarrow x_i^*$ on a set of variables Y disjoint of $(X_i \text{Cpa}(X_i))$, the marginal probability $p(y \mid x_i^*)$ obtains by summing this density over all variables except $Y \cup X_i$ (“adjustment for direct causes” - see Pearl, 2000, p. 73). More generally, the identification of the effects of one or more interventions on a causal system “typically involves a sequence of operations that gradually transform expressions involving intervention probabilities to expressions involving ordinary conditional probabilities, the latter being in principle accessible by empirical observation” (Lauritzen, 2001, p. 83; see also Pearl, 1995, 2000 Chap. 3).

4.5. Directed Acyclic Graphs and INUS Condition

One could wonder whether the factorial causation models examined in Section 4, based on INUS condition or some of its generalizations, may be more or less compatible with the DAG approach, so as one could choose either approach, exploiting the easier or more meaningful features in each particular application.

As Aickin (2002, pp. 205-210) clearly states, the two approaches are remarkably different, and perhaps the effort towards a possible reconciliation is worthless. The fact that Aickin’s theory connects events and not variables is not, by itself, a hindrance: it is always possible to interpret the events $A, B, \dots$ as dichotomous variables (indicators) taking values 0 and 1. With reference to the INUS condition (16), and assuming for a specific situation that A is independent of (B, C) , and B has a causal influence - and time precedence - on C , the following DAG could hold:



with the following specification of probabilities:

$$p(d \mid a, b, c) = 1 \quad \text{if } a = 1 \text{ (any } b \text{ and } c) \\ \text{or if } b = 1, c = 1 \text{ (any } a).$$

This example may suggest that actually most DAGs do not satisfy a property like the one above, i.e. that the “effect” is bound to occur at least when all the “explanatory factors” do occur. It appears that “few if any DAGs specify the causal field to the point that causal relations are deterministic” (Aickin, 2002, p. 210).

5. Confounding, Randomization, Counterfactuals

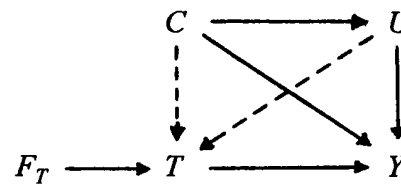
5.1. Countering Confounding by Randomization

A comfortable way of managing interventions in a DAG is by means of *augmented* DAGs, which are constructed by simply adjoining - to one or more nodes - an *intervention variable*; practically, if X is a variable in a DAG, the intervention variable F_X is introduced as in Figure 3; F_X can take all the values of X , meaning that $F_X = x$ indicates the intervention $X \leftarrow x$; instead, $F_X = \emptyset$ (empty set) means that X is allowed to assume values as produced by the natural mechanism which is reflected in the joint probability of all

nodes.

Figure 3 displays the causal relationships existing between a *treatment* T , a *response* Y , observed *covariates* C , an *unobserved* confounding variable (or *confounder*) U , and an intervention variable F_T . The dashed lines from C to T and from U to T can be turned to solid lines, or else can be deleted. A solid line from C to T means an influence of the observed covariates on treatment T ; a solid line from U to T means that the assignment of treatment is affected by the unobserved covariate U , which in this case becomes a confounder. The variable U is a confounder because it affects both the treatment T and the response Y ; if it could be observed, responses could be adjusted for variation in U , by partitioning the reference population of units into strata that are homogeneous with respect to U , thus allowing comparisons of treatments and responses *given* U . “The concept of confounding is a central one in modern epidemiology. Although confounding occurs in experimental research, it is a considerably more important issue in nonexperimental research ... A large portion of epidemiologic methods are concerned with avoiding or adjusting (controlling) for confounding” (Rothman and Greenland, 1998, pp. 120, 62).

Figure 3: A causal framework for randomized and nonrandomized studies



The easiest - when feasible - and most adopted way of cutting the link between U and T is by means of *randomization* in the assignment of treatments to subjects, meaning that such assignment is made by a randomized mechanism, whose functioning is independent of any other factor. Actually, such a definition is properly related to “simple randomization”; but this same kind of assignment can be made to depend on known covariates, giving rise to “stratified randomization” (in this case there should be a solid line between C and T in Figure 3 - cf. Edwards, 2000, p. 233).

Thus, if the arrow from U to T is deleted, it is easy to control (algebraically, by means of a joint probability density consistent with the graph, or exploiting the graphical properties of DAGs - see e.g. Lauritzen, 2001), the conditional independence relation $T \perp U \mid C, F_T$, which is assumed by Dawid (2002, p. 177) as the operational description of the “randomization condition”; when this condition is satisfied, it is possible to assess the *treatment effect*, i.e. the influence on the response Y of the treatment, given the observed covariates. Such an assessment can be more or less analytical, depending on the available information concerning the joint probability density, or some of its essential features.

As a rule, only one treatment can be applied to each unit or subject in an experimental setup; this is because a “unit” must be appreciated in all its features, that can be affected - or interact - with the treatment. A natural consequence is that only in exceptional situations can *the same unit* be given two - or more - different treatments in successive times; thus we shall assume that only one treatment is given to each unit. For the sake of simplicity, but practically without losing in generality, we shall assume that only one in two treatments t and c (control) is given to each unit.

A classical example is taken from Greenwood and Yule (quoted in Suppes, 1970, p.

12); the following table is concerned with the efficacy of inoculation against cholera, as tested from 818 subjects, 279 inoculated and 539 not-inoculated.

	Not attacked	Attacked	Totals
Inoculated	276	3	279
Not-inoculated	473	66	539
Totals	749	69	818

From usual summaries of like data, the efficacy of inoculation seems quite ascertained (letting aside sampling variability); for example, the proportion of attacked within inoculated is $3/279 = 1,08\%$, while the proportion rises to $66/539 = 12,24\%$ for those who were not inoculated. However, these are not the data one would dispose of; such data should be, as an example, as those appearing in the following table.

	Not attacked	Attacked	Totals
Inoculated	806	12	818
Not-inoculated	695	123	818

To sum up, one would dispose of the response Y (not-attacked, attacked) *for each of the treatments*; but this is obviously impossible. Such impossibility cannot exempt us from the search (i.e. estimation) of the proportion of responses related to each treatment *for each unit*, as this is just the object of the investigation. The answer to this search is positive when the treatment assignment is effected by means of randomization.

Let $Y_t(u)$ and $Y_c(u)$ be the responses to the two treatments on the unit u , and let us assume that we are interested in the average of the causal effects, given by the difference

$$C(u) = Y_t(u) - Y_c(u) \quad (30)$$

for all units in a given population (finite or virtually infinite).

Although seemingly reductive, we shall make the simplifying assumption that there is a unique value $Y_i(u)$ ($i = t, c$) corresponding to unit u and treatment i (so-called “stable unit-treatment value assumption, or SUTVA - see Rubin (1980), Rosenbaum and Rubin (1983)); of course, this assumption is quite justified if the “unit” is described by all relevant features and circumstances present when treatment is implemented, and if treatments are the same - in all relevant characteristics - for all units; also, the response $Y_i(u)$ must not be affected by the treatments and responses concerned with other units.

If we dispose of a random sample of size n from the reference population, we can make the assignment of treatment t to n_t units and of treatment c to n_c units ($n_t + n_c = n$). A random variable S is defined, which takes the value t on unit u if such unit is assigned treatment t , and the value c if the treatment is c . From Suppes (1970) and Holland (1986) we may call *prima facie causal effect* the following quantity

$$C_{PF} = E_S[Y_t(u) | S(u) = t] - E_S[Y_c(u) | S(u) = c] \quad (31)$$

where E_S is the average operator within the sample values indicated in square brackets; in words, C_{PF} is the difference in the means of the n_t responses that have received treatment t , and the n_c responses that have received treatment c . C_{PF} is obviously computed in order to get an *estimate* of the sample mean

$$\bar{C} = E_S[Y_t(u)] - E_S[Y_c(u)] \quad (32)$$

which is obviously counterfactual, as we do not have knowledge of $Y_t(u)$ and $Y_c(u)$ for *all* the units in the sample. Such estimate C_{PF} is easily seen to be generally biased (Edwards, 2000, p. 227).

A simple way to overcome this difficulty, and getting an unbiased estimate, lies in assigning treatments by a random mechanism. The most natural randomization, which consists in choosing a simple random sample (of size n_t for treatment t and size $n_c = n - n_t$ for treatment c) from the finite population of n units, gives rise to estimates

$$E_S[Y_t(u) | S(u) = t] \quad \text{unbiased for } E_S[Y_t(u)] \quad (33)$$

$$E_S[Y_c(u) | S(u) = c] \quad \text{unbiased for } E_S[Y_c(u)] \quad (34)$$

as they are sample means of a random sample; thus the difference between these estimates is unbiased too with respect to C (mean of C values over the whole population).

Of course, this last conclusion can be supported only when we dispose of a random sample from the reference population, or - anyway - when the representativeness of the sample may be duly established; as in many cases samples are not obtained by random drawing, to be honest one should at least justify *which* population is compatible with the given sample.

Likewise, for the usual case of matched treatment-control pairs, the sample mean of the *observed* causal effects may be written (using r_1 and r_2 for the first and second unit of a pair, ordered by the flip of a coin), letting $S(r)$ a Bernoulli variable taking values 0 and 1 with probability 0.5:

$$\frac{n}{2} \sum_{r=1}^{n/2} \{ [Y_t(r_1) - Y_c(r_2)]S(r) + [Y_t(r_2) - Y_c(r_1)](1 - S(r)) \} \quad (35)$$

By averaging this estimator one obtains

$$E_S[C(u)] = \frac{1}{n} \sum_{u=1}^n [Y_t(u) - Y_c(u)] = \frac{1}{n} \sum_{u=1}^n C(u) \quad (36)$$

i.e. the sample mean of all causal effects, which in its turn is unbiased with respect to C .

It must be added that the practical implementation of randomization is by no means simple or automatic; everything must be possibly controlled so that the treatments can really be the same for every unit, so that only the effects of treatments - and not of other concomitant factors - can eventually be measured. This is clearly appreciated, for example, in the medical *double-blind* experiments: “for the benefits of random assignment to be realized fully, neither the researcher administering the cure and the placebo nor the subjects

in the control and the treatment groups can know who is in what group and which tablet is which” (Zeisel and Kaye, 1997, p. 3).

5.2. Pseudo-randomization in Observational Studies

Most scientific investigations are made outside laboratories, and in any case are not subjected to the precise requirements of the controlled randomized experiments: they are called *observational studies*, and are subjected to all sorts of possible errors and shortcomings. It is with respect to these studies that practically all philosophical thoughts have elaborated canons, or criteria - such as those of J.S. Mill, and those elaborated by A.B. Hill (1965 - see also Holland, 1986) mainly for epidemiologic investigations. Needless to say, all these proposals have been heavily criticized, sometimes with a constructive purpose (see the contributions in the volume edited by K.J. Rothman, 1988).

A simple and clever proposal to circumvent such difficulties was made by Rubin (1974, 1977, 1978), and later elaborated by himself and other scholars (Rosenbaum and Rubin, 1983; Rosenbaum, 1984; Holland, 1986). The starting point is again the *prima facie* causal effect (31), however referred to a specification $X = x$ of the vector X of covariates:

$$C_{PF} = E_S[Y_t(u) | X = x, S(u) = t] - E_S[Y_c(u) | X = x, S(u) = c]. \quad (37)$$

We now posit the assumption of a *strongly ignorable treatment assignment* (so-called SITA condition), i.e.

$$(Y_t, Y_c) \perp S | X \quad \text{being } 0 < P(S = t | X = x) < 1 \quad (38)$$

which obviously holds in stratified randomization; in other words, we are making the strong assumption that we have been so clever - and fortunate - as to find out covariates X so that, in each stratum specified by $X = x$, the assignment mechanism S (of the treatments) is independent of the responses, which is the same as saying that there is a “natural” randomization within the strata.

Analogously with the comment made for (33) and (34), the estimate

$$E_S[Y_i(u) | X = x, S(u) = i] \quad \text{is unbiased for } E_S[Y_i(u) | X = x] \quad (39)$$

for $i = t, c$; moreover, such estimate is unbiased for $E_S[Y_i | X = x]$ in the population. As a consequence, the weighted mean of these estimates, with weights equal to $P(X = x)$ for each stratum, is unbiased for $E(Y_i)$, while the weighted mean with weights equal to the sample proportions is unbiased for $E_S[Y_i(u)]$. Of course, both estimates coincide when the available random (?) sample is actually a *proportional* stratified sample.

Something similar can happen if - maintaining that the treatment assignment is ignorable given the covariates - instead of conditioning on $X = x$, one conditions on a “balancing score” (Rosenbaum and Rubin, 1983), whose coarsest version is the function

$$e(x) = P(S = t | X = x),$$

i.e. the probability of receiving a treatment given the observed covariates; such a function is called “propensity score”.

Rosenbaum (1984), while admitting the great difficulty of making inferences from observational studies (“Even after adjustments have been made for observed covariates ... estimates of treatment effects can still be biased by imbalances in unobserved covariates”, p. 41; “... in observational studies, ignorable treatment assignment is but a tenuous

assumption”, p. 43), is nevertheless persuaded that strongly ignorable treatment assignments can be recognized, hence allowing the computation of unbiased and consistent estimates. To this end, he suggests an approach to testing the SITA hypothesis, by exploiting relevant consequences (equalities and inequalities) ensuing from the causal mechanism - or accepted theory - that governs the specific treatment-response setup.

As most discoveries in the social and biological sciences have been effected by means of observational studies - and in most cases a controlled randomized experiment is neither practical nor ethically admissible - every contribution towards gaining rigour for inference in observational studies must be welcomed. Therefore Rubin’s proposal deserves to be attentively scrutinized by statisticians and empirical scientists, in order to yield correct procedures outside randomized experiments. In any case, the application is neither simple nor can be taken for granted. The SITA condition (38) is really strong, and we cannot assume that nature behaves as a professional experimenter in his/her laboratory, although this can sometimes occur. Even the strata weights, and the propensity scores, can usually be unknown, and hardly estimable. To all this, an essential ingredient must be added, in order to get the desired inference for the reference population: the available sample - altogether and within strata - must be a random sample, anyhow a “representative” sample, a condition which is too often neglected, perhaps on purpose, because it is usually impossible to be ascertained.

5.3. The Role of Counterfactuals

In the above subsections we have hinted at the counterfactual nature of some entities; perhaps a few words could be expedient. A counterfactual (statement, or conditional) is an assertion of the form “if X had been the case, then Y would have happened”, made when it is known to be false that X is the case (Dawid, 2000, p. 408). For example, if I had arrived at the station five minutes earlier, I should have not missed my train (of course I missed it). A poetical version of this aseptic definition is given by the Italian poet Guido Gozzano (from *Cocotte*):

Non amo che le rose
che non colsi. Non amo che le cose
che potevano essere e non sono
state ...

Just “the things that could have been and never happened” is a good definition of counterfactuals. Counterfactual statements, as well as subjunctive conditionals (of the form “if X would be the case, then Y would happen as well”) share a distinctive place in the theory of scientific explanation (Goodman, 1965-1983, Chap. 1; Hempel, 1966, Chap. 5); it is widely acknowledged by the philosophers of science that counterfactual and subjunctive conditionals can feature scientific assertions, as contrasted to non-scientific assertions, which are unable to support such conditionals. All science is essentially counterfactual; otherwise it could be only *description* of known facts - as accurate as it can be. A scientific law - be it deterministic ($a \Rightarrow b$) or probabilistic ($a \Rightarrow b$ with probability p) - bears no reference to precise historical facts, but applies to all facts of a certain kind.

Concerning the subject of this subsection, some misunderstandings have arisen with respect to the *nominal* status of certain quantities, and to the inferential use of these same quantities. All comes from the (customary) complementary nature of the responses $Y_t(u)$ and $Y_c(u)$ to treatments t and c , as only one of them can be observed, the other being

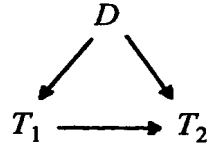
counterfactual once a particular treatment has been assigned to unit u . Thus, first of all, it must be said that $Y_f(u)$ and $Y_c(u)$ have nothing counterfactual in themselves *before* the treatment assignment, although it does not seem really mistaken to speak of a *counterfactual framework* (Edwards, p. 226), thinking of what can happen *after* the treatment has been assigned. For this reason, it seems preferable to share Rubin's proposal (2000, p. 435) to call $Y_f(u)$ and $Y_c(u)$ *potential outcomes*, as they really are before treatment assignment. Moreover, only the labels of potential outcomes and the observations ensuing from given treatments - as can be seen from (32)-(39) - are functional in these formulas; of course, how could one use unknown values for inference purposes? No problem - of course - bears on unknown or unknowable quantities as the *object* of inference, which is the only real problem in science and operational research.

What then? There remain questions that cannot be asked without some knowledge of counterfactual responses. One important question, for example, concerns the variability of the estimator of the population causal effect C , given by the difference between (33) and (34); writing the variance of this estimator requires knowledge of the correlation between the two responses for each unit; however, as only one of the responses can be observed, the correlation itself is counterfactual (it could be estimated if we had observed both responses for each sample unit). If something can be done in these cases, you are required to exploit any information outside the experimental framework and its outcomes, possibly achieving inequalities for the parameters of interest; anyway, abandoning the counterfactual framework can sometimes be feasible and rewarding, but no general achievement seems to have been available so far (see Dawid, 2000, and subsequent discussion).

6. Conditional Independence and Causal Discovery

Some closing remarks are addressed to the assessment of some basic questions, concerned with the construction of causal models, and with their validation by means of data. The first question is related to the connected topics of the Causal Markov Condition and implied conditional independencies in the model (cf. Section 4). Given a graph with n vertices (variables), it is possible - in principle - that every vertex is connected with every other vertex in the graph: in this case, the graph is called *complete*, in which every pair of its vertices are *adjacent* (i.e. there is an edge connecting the given vertices). Note that the connecting edges, if endowed with arrows, would easily yield a non-DAG graph, as one or more loops could arise. A complete graph, with more than three variables, is not very interesting: it means that all the variables have some influence on each other. In science, and mostly in operational research, one aims at discovering simple structures, where some variables act on some others, these - in their turn - act on some others etc. All this brings about a number of conditional independencies, many of them "justified" by the Humean condition of "no action at a distance".

Quite to the contrary, this condition of contiguity in space and time is wholly devoided of any empirical meaning. This is easily seen. For example, let $D = (D_1, D_2)$ be a variable "disease", which in a certain instance can take two values D_1 and D_2 (two distinct diseases), and T_1 and T_2 be the temperature taken at successive times t_1 and t_2 . It is evident that the model $D \rightarrow T_1 \rightarrow T_2$ is wholly mistaken, as the temperature at time t_2 does depend on the previous temperature T_1 , but of course it also depends on the specific disease which is present. Thus the correct model should be the following:



With reference to most biological phenomena, another observation must be added. Some researches, especially in the philosophical domain, sustain that the action of human beings has a *finalistic* interpretation, so that a specific action *today* is mostly - if not exclusively - *explained* or *caused* by the goal-setting established by the individual. In a sense, the (desired) effect exerts an influence on the cause - i.e. what is arranged for achieving the effect. Thus, it seems that the Humean temporal precedence of the cause with respect to the effect may be reversed. We shall not delve into this puzzling “interchange” between cause and effect, but this same cue will be exploited for the related sense of an “unwanted” or “unconscious” *project*, which is imprinted in the DNA of any living being.

As an example, let us consider the case of broad-leaved trees: their leaves fall as winter approaches - at our latitude; the *apparent cause* of falling leaves, in the Humean sense, is just the succession of seasons and drop in temperature. But the *real cause* - by answering the question *why*, and not only *how* - is contained in the *project* existing inside each cell of the tree: their ancestors, in millions of years, have learnt that falling leaves is functional to get over winter, unharmed and ready to flourish again in the following spring; the project is towards survival; the content in water of the leaves would not permit to endure cold winters. If we should follow the life of a single tree, the DNA project could be (in the right?) assumed in the background. But if we follow the life of several trees, possibly of diverse species, the above project - with distinctive although similar features for each tree - must be placed into the model as the principal and *constant* (in time) cause of everything happening. As for the example of disease and temperature, “changes” contiguous in time and space are unable to explain what is going on.

Anyway, looking for simple structures in multivariate models - i.e. with many conditional independencies - may be rewarding, provided the approximation to reality is reasonable, and no forcing is exerted; otherwise the model would become impracticable. As Pearl (2000, p. 44) writes, “the Markov assumption is more a convention than an assumption”; thus the research worker must be careful not to disallow the actual structure of the real phenomena.

That Markov assumption must not be taken too literally -or seriously -becomes evident if we think of the consistency of all the probabilities involved. For example, let us consider a model with three variables A , B and D , and let us assume that a “common cause” - or conjunctive fork - scheme holds, so that D is the common cause of A and B , so that $(A \mid D) \perp (B \mid D)$. The probabilities for the marginal (A, B) are obtained from the joint distribution of (A, B, D) , under the above conditional independence assumption, by

$$P(A, B) = \sum_D P(A, B \mid D)P(D) = \sum_D P(A \mid D)P(B \mid D)P(D).$$

In the simple case of a dichotomous $D = \{D_1, D_2\}$, $A = \{A_1, \dots, A_h\}$, $B = \{B_1, \dots, B_k\}$, such an expression becomes:

$$P(A_r, B_s) = P(A_r \mid D_1)P(B_s \mid D_1)P(D_1) + P(A_r \mid D_2)P(B_s \mid D_2)P(D_2)$$

for $r = 1, \dots, h$ and $s = 1, \dots, k$. Given the distribution for the cause D and the conditional probabilities $P(A | D)$ and $P(B | D)$, the distribution for the marginal (A, B) is obviously and immediately obtained. However, if we change our viewpoint, starting from the “observation” of the contingency table for (A, B) , and look for a possible probability structure consistent with the above conditional independence, some serious problems may arise: we have to look for solutions in the above non-linear system. Taking account of the sum conditions on $P(D_i)$, $P(A_r | D_i)$, $P(B_s | D_i)$, and also that one of the equations is a linear combination of the remaining ones, we have $hk - 1$ equations in $2(h + k - 1)$ unknowns, which become $2(h + k - 2)$ in the case of knowledge of D distribution - which in itself has nothing to do with conditional independencies. Thus it can easily happen, for realistic values of h and k (relatively small contingency tables), that no solution exists - i.e. the contingency table is inconsistent with the assumption of conditional independence - or else that the set of solutions is constituted of just one point, or the points of a unidimensional curve in a multidimensional space.

These observations have a direct impact on causal discovery algorithms, which aim at identifying causal relationships on the basis of sample data; some of these algorithms have been implemented in the program TETRAD (several versions from 1987 - see Spirtes *et al.*, 2000). Letting aside the evergreen requirement that the relevant variables - for a given problem - have been included in the model, let us concentrate on two related questions: (a) point hypotheses, and (b) consistency of statistical tests.

Starting with a complete graph (all nodes connected, with arrow directions determined by time precedence, when available), the model becomes more and more *causal* by successively deleting edges, on the basis of recognition of conditional independencies; “in the discrete case the procedure performs tests of conditional independence and in the linear continuous case tests for vanishing partial correlations” (Spirtes *et al.*, 2000, p. 82), which in the normal case are equivalent to conditional independence. Such a procedure suffers from some well known but often overlooked weaknesses. Perhaps the more apparent descends from the *multiple* test procedure itself: “In repeated tests of independence hypotheses in a sample, each using the same significance level, the probability that *some* true hypothesis will be rejected is not given by the significance level; depending on the number of hypotheses and the sample size, that probability may in fact be much higher than the significance level ... “ (Spirtes *et al.*, 2000, p. 96).

However, another important weakness must be identified in the very nature of the hypotheses tested, namely (conditional) independence between variables. The independence concept, so useful as it is - must nevertheless be given its essentially *theoretical* nature: “it is hard to find things that are truly independent” (Lindley, 1979, p. 15). But also in this abstract position, the independence concept needs some careful attention. As underlined above, the independence hypothesis is a *point null hypothesis*, with a clear theoretical status, but hardly tenable when confronted with the real world: with practical certainty, we could declare it false from the outset. As such, it looks like many other null hypotheses, for example: (a) the mean of a normal variable is $\mu = 2$ (a precise real number), (b) the parent distribution of the sample is normal, etc. Fortunately enough, in the Neyman-Pearson approach to testing hypotheses such statements are not taken at nominal value, although it would be preferable to be more explicit, for example by fixing an interval for the mean like $(1.98, 2.02)$ instead of $2,000\dots$ (Frosini, 2001, p. 374). With this natural and sensible interpretation, the desirable property of consistency does not yield - in the limit - such embarrassing results as those pointed out by the Bayesian in the discussion of so-called

Lindley paradox (Berger, 1985, pp.148-156); as shown by Frosini (2003), no puzzling consequence is obtained if one avoids assigning a finite probability to a point null hypothesis, which is absurd because such a hypothesis is (practically) false.

In conclusion, the independence hypothesis could be *operationally* maintained if not taken too seriously, exactly as for every other point hypothesis. And in testing such hypotheses one must be careful about the effect of the consistency property (of tests): with standard significance levels, the null hypothesis is regularly accepted when the sample size is small, and regularly rejected when the sample size is large. Perhaps these simple facts are responsible for the contrasting outcomes of two interesting examples discussed by Spirtes *et al.* (2000, pp. 103-108; original data from Bentler and Peeler, 1979, and Blau and Duncan, 1967), concerned with factors for the female orgasm and the American occupational structure. An apparently strong causal structure in the first case and an almost complete graph in the second case could be the outcome - at least partially - of a sample size of 281 female university undergraduates with seven variables in the first case, and a sample size of 20,700 workers with five variables in the second case.

References

- Aickin M. (2002) *Causal Analysis in Biomedicine and Epidemiology*, Marcel Dekker, New York.
- Anderson J. (1938) The problem of causality, *Australasian J. of Psychology and Philosophy*, 16, 127-142.
- Andersson S.A., Madigan D., Perlman M.D. (1997) A characterization of Markov equivalence classes for acyclic digraphs, *Ann. Statist.*, 25, 505-541.
- Aristotle (1928) *Analytica Posteriora*, in : The Works of Aristotle, Ross W.D (Ed.), vol. 1, Clarendon Press, Oxford.
- Aristotle (1930) *Physica*, in : The Works of Aristotle, Ross W.D (Ed.), vol. II, Clarendon Press, Oxford.
- Bentler P. and Peeler W. (1979) Models of female orgasm, *Archives of Sexual Behavior*, 8, 405-423.
- Berger J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.
- Blau P. and Duncan O. (1967) *The American Occupational Structure*, Wiley, New York.
- Bickel P.J., Hammel E.A., O'Connell J.W. (1977) Sex bias in graduate admissions: Data from Berkeley, in: *Statistics and Public Policy*, Fairley W.B. and Mosteller F. (Eds.), Addison-Wesley, Reading, MA.
- Carnap R. (1950, 1962) *Logical Foundations of Probability*, Second Edition 1962, University of Chicago Press, Chicago.
- Copi I.M., Cohen C. (1998) *Logic*, Tenth edition, Prentice Hall, Upper Saddle River, NJ.
- Dawid A.P. (1979) Conditional independence in statistical theory, *J. Roy. Statist. Soc. B*, 41, 1-31.
- Dawid A.P. (2000) Causal inference without counterfactuals (with Discussion), *J. Am. Statist. Assoc.*, 95, 407-448.
- Dawid A.P. (2002) Influence diagrams for causal modelling and inference, *International Statistical Review*, 70, 161-189.
- Edwards D. (2000) *Introduction to Graphical Modelling*, Second Edition, Springer, New York.
- Eells E. (1991) *Probabilistic Causality*, Cambridge University Press, Cambridge.

- Faliva M. (1992) Recursiveness vs. interdependence in econometric models: a comprehensive analysis for the linear case, *J. Ital. Statist. Soc.*, 1, 335-357.
- Faliva M. and Zoia M.G. (1994) Detecting and testing causality in linear econometric models, *J. Ital. Statist. Soc.*, 3, 61-76.
- Fisher R.A. (1959) *Smoking: the Cancer Controversy*, Oliver and Boyd, Edinburgh.
- Frosini B.V. (1999) Conditioning, information, and frequentist properties, *Statistica Applicata*, 11, 165-184.
- Frosini B.V. (2001) *Metodi statistici*, Carocci, Roma.
- Frosini B.V. (2002) *Le prove statistiche nel processo civile e nel processo penale*, Giuffrè, Milano.
- Frosini B.V. (2003) On Neyman-Pearson theory: Information content of an experiment and a fancy paradox, *Università Cattolica di Milano, Istituto di Statistica, Serie E.P.* N. 120.
- Frydenberg M. (1990) The chain graph Markov property, *Scand. J. Statist.*, 17, 333-353.
- Götze A. (1947) *Old Babylonian Omen Texts*, Yale University Press, New Haven.
- Goodman L.A. (1973) The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach, *Biometrika*, 60, 179-192.
- Goodman N. (1983) *Fact, Fiction, and Forecast*, Fourth Edition, Harvard University Press, Cambridge, MA.
- Hempel C.G. (1962) Deductive-Nomological vs Statistical Explanation. in: *Minnesota Studies in the Philosophy of Science*, Feigl H. & Maxwell G. (Eds.), University of Minnesota Press, Minneapolis.
- Hempel C.G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, Free Press, New York.
- Hempel C.G., Oppenheim P. (1948) Studies in the logic of explanation, *Philosophy of Science*, 15, 135-175.
- Hill A.B. (1965) The environment and disease; association or causation, *Proceedings of the Royal Society of Medicine*, 58, 295-300.
- Holland P.W. (1986) Statistics and causal inference, *J. Am. Statist. Assoc.*, 81, 945-970.
- Hume D. (1888) *A Treatise of Human Nature*, Clarendon Press, Oxford.
- Jeffrey R.C. (1969) Statistical explanation vs. statistical inference, in: *Essays in Honor of Carl G. Hempel*, Rescher N. (Ed.), D. Reidel, Dordrecht.
- Laplace P.S. (1840) *Essai philosophique sur les probabilités*, Septième édition, Société Belge de Librairie, Bruxelles.
- Lauritzen S.L. (1996) *Graphical Models*, Clarendon Press, Oxford.
- Lauritzen S.L. (2001) Causal inference from graphical models, in *Complex Stochastic Systems*, Barndorff-Nielsen O.E. (Ed.), Chapman and Hall, London, 63-107.
- Lauritzen S.L., Dawid A.P., Larsen B.N., Leimer H.-G. (1990) Independence properties of directed Markov fields, *Networks*, 20, 491-505.
- Lindley D.V. (1979) Discussion of Dawid's paper, *J. Roy. Statist. Soc. B*, 41, 15-16.
- Mackie J.L. (1965) On causes and conditions, *American Philosophical Quarterly*, 245-255 and 261-264.
- Mackie J.L. (1974) *The Cement of the Universe: A Study of Causation*, Oxford University Press, Oxford.
- Mill J.S. (1843) *A System of Logic*, J.V. Parker, London.
- Pearl J. (1995) Causal diagrams for empirical research, *Biometrika*, 82, 669-710.
- Pearl J. (2000) *Causality. Models, Reasoning, and Inference*. Cambridge University Press,

- Cambridge.
- Pearson K. (1900) *The Grammar of Science*, Second Edition, Adam and Charles Black, London.
- Reichenbach (1956) *The Direction of Time*, University of California Press, Berkeley.
- Rosenbaum P.R. (1984) From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment, *J. Am. Statist. Assoc.*, 79, 41-48.
- Rosenbaum P.R., Rubin D.B. (1983) The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, 41-55.
- Rothman K.J. (1976) Causes, *American J. of Epidemiology*, 104, 587-592.
- Rothman K.J. (Ed.) (1988) *Causal Inference*, Epidemiology Resources, Chestnut Hill, MA.
- Rothman K.J., Greenland S. (1998) *Modern Epidemiology*, Lippincott-Raven, Philadelphia.
- Rubin D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies, *J. of Educational Psychology*, 66, 688-701.
- Rubin D.B. (1977) Assignment to treatment group on the basis of a covariate, *J. of Educational Statistics*, 2, 1-26.
- Rubin D.B. (1978) Bayesian inference for causal effects: the role of randomization, *Ann. Statist.*, 6, 34-58.
- Rubin D.B. (1980) Discussion of Basu's paper. *J. Am. Statist. Assoc.*, 75, 591-593.
- Rubin D.B. (2000) Comment on Dawid's paper, *J. Am. Statist. Assoc.*, 95, 435-438.
- Russell B. (1913) On the notion of cause, *Proceedings of the Aristotelian Society*, 13, 1-26.
- Salmon W.C. (1965) The status of prior probabilities in statistical explanation, *Philosophy of Science*, 32, 137-146.
- Salmon W.C. (1978) Why ask "Why"?: An inquiry concerning scientific explanation. *Proceedings and Addresses of the American Philosophical Association*, vol. 5, no. 6, 683-705
- Salmon W.C. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton.
- Scriven M. (1959) Explanation and prediction in evolutionary theory, *Science*, 130, 477-482.
- Simpson C. (1951) The interpretation of interaction in contingency tables, *J. Roy. Statist. Soc. B*, 13, 238-241.
- Spirtes P., Glymour C., Scheines R. (2000) *Causation, Prediction, and Search*, The MIT Press, Cambridge, MA.
- Suppes P. (1970) *A Probabilistic Theory of Causality*, North Holland, Amsterdam.
- Verma T., Pearl J. (1990) Equivalence and synthesis of causal models, in: *Uncertainty in Artificial Intelligence*, Proceedings of the Sixth Conference, Morgan Kaufman, San Francisco, 220-227.
- Wermuth N., Lauritzen S.L. (1983) Graphical and recursive models for contingency tables, *Biometrika*, 70, 537-552.
- Wright S. (1921) Correlation and causation, *J. Agric. Res.*, 20, 557-585.
- Wright S. (1934) The method of path coefficients, *Ann. Math. Statist.*, 5, 161-215.
- Yule G.U. (1903) Notes on the theory of association of attributes in Statistics, *Biometrika*, 2, 121-134.
- Zeisel H., Kaye D. (1997), *Prove it with Figures*, Springer, New York.