Journal of Statistical Planning and Inference 25 (1990) 261–278
North-Holland

CAUSALITY AND STATISTICS*

A.P. DEMPSTER

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Received October 1987; revised manuscript received March 1989
Recommended by K. Hinkelmann

Abstract: Many aspects of statistical design, modelling, and inference have close and important connections with causal thinking. These are analyzed in the paper against a philosophical background that regards formal mathematical models as having dual interpretations, reflecting both objectivist reality and subjectivist rationality. The latter aspect weakens the need for an objective theory of probabilistic causation, and suggests that a traditional image of causes as deterministic mechanisms should remain primary. It is argued that such causes should guide much preformal thinking about what to include in formal statistical models, especially of dynamic phenomena. The statistical measurement of causal effects is facilitated by good statistical design, incuding randomization where feasible, and requires other methodologies for controlling and assessing uncertainties, for example in model construction and inference. Illustrative examples include case studies where the problem is to assess retrospectively the causes of observed events and where the task is to assess future risks from controllable factors.

Key words: Causal inference; causal effects; subjectivity; objectivity; randomization; experimentation.

1. Introduction

Why should statisticians analyze causality? The first reason is practical. Causal thinking is deeply embedded in scientific understanding of the problems of applied statistics. Would a reduction of alcohol consumption by women eliminate or postpone some cases of breast cancer? Were discriminatory employment practices a cause of the male–female differentials at the Harris Bank that led a decade later to a 14 million dollar settlement (New York Times, Jan. 11, 1989)? Was the death of the patient analyzed by Lane (1989) caused by an adverse reaction to the drug amodiaquine? Will the build-up of cardon dioxide in the earth's atmosphere cause catastrophic climate changes in the next hundred years? Statistical data collection and analysis constitute much of the empirical basis of credible attempts to answer such questions. As the examples show, familiarity with the language and ideas of causality is essential to full participation in inference and decision-making.

* Research supported in part by NSF grant DMS-88-07085.

Statisticians typically bring a range of qualifications to the task of analyzing and explaining causation that may in important ways improve upon those of philosophers. Statisticians participate directly and indirectly in scientific developments in many fields where statistical methodology is applied, so are well placed to develop a philosophy of science, including statistical science, that accords with the realities of practice. Wisdom on the possibilities and pitfalls of passing from empirical relations to causal inference has been clearly articulated by statistical writers (e.g., Cochran, 1965). Many statisticians have strong mathematical training, and hence are equipped to appreciate the power of abstract models to represent and shape scientific perceptions and analyses. Finally, applied statisticians frequently confront scientific uncertainty, and in particular must deal implicitly and explicitly with competing and interacting subjectivist and stochastic conceptions of probability. Inferences about deterministic causal phenomena are often uncertain inferences that must be viewed, whether formally or informally, in subjectivist terms, while the stochastic models that permeate modern science can be interpreted as objective causal random mechanisms (Good, 1961/2, 1988, or Suppes, 1970, 1988). The examples mentioned above, and further discussed below, document the pervasive interactions of uncertainty and causal analysis, yet many books on the philosophy of science virtually ignore probabilistic uncertainty.

I accept as fundamental the common sense view that causes are primitive elements of scientific thinking from which informal and then formal understanding of dynamic phenomena develops. The central idea of causation is the mental image of a causal mechanism acting in a deterministic physical way that necessarily and regularly produces a defined subsequent effect. It is implicit that had the mechanism been absent, or different, the effect might also have been different (Holland, 1986). An illustrative quote from Fisher (1918) on the 'causes of human variability' is:

> If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and in that case he would have been shorter.

Many writers are impatient with the common sense view, labelling it the 'metaphysical idea of Causation' (Herschel, 1851 quoted by Porter, 1986) or 'prescientific' (Russell, 1948). Pearson (1911) criticizes both 'spiritualist' and 'materialist' conceptions:

> Force as cause of motion is exactly on the same footing as a tree-god as cause of growth – both are but names that hide our ignorance of the why in the routine of our perceptions. The necessity in a law of nature has not the logical must of a geometrical theorem, nor the categorical must of a human law-giver; it is merely our experience of a routine, whose stages have neither logical nor volitional order.

These writers imply that abstract representation has rendered causal analysis un-

important. I agree that causal thinking is largely preformal, but formal analyses are never firm, especially in statistics, so must remain in close touch with their preformal roots. Hence these roots deserve to be considered important aspects of the process.

The spectacular successes of mathematical models and mathematical expressions of scientific laws brought in their wake the idea that abstract representations should replace mechanistic causal images as the central elements of scientific explanation. But, whereas a causal mechanism can be seen as explaining a small piece of a scientific mosaic, a formal mathematical model describes its own complete small world. And, whereas available procedures for manipulating, observing, recording, and analyzing indicate that the external world is complex without limit, mathematical explanation simplifies understanding at the cost of limiting discussion to a universe that is a priori unrealistic. The problem of small worlds is vexing for mathematical scientists (Savage, 1954, Shafer, 1986). Mathematical modelling is fundamentally important, but does not itself supplant the notion of causal mechanism. The latter feeds the former in an endless quest to catch up with real world complexity.

Causal analyses are guides to higher understanding. Although mathematical models provide formal representations whose precision and efficiency far surpass the reach of nonmathematical language, they are understanding-neutral in the sense of not explaining how and why their features capture key aspects of reality. Causal thinking is a fundamental tool in the development of balanced representations that neither omit essentials nor obscure with unnecessary complexity. An account of causality requires an explanation of how mathematical models relate both to real world processes under study and to the memory and reasoning capabilities of the community of scientists who perceive and use such models. The following discussion explores these themes.

2. Philosophical context

Because views on causality are necessarily linked in complex ways with views on many other philosophical issues, causation cannot be studied in isolation from the network of basic concepts and positions that make up a philosophical viewpoint. Working scientists, including statisticians, typically operate in modes imparted by their teachers and textbooks, which in turn depend on philosophical positions that may be implicit, unexamined, and naive. One can agree with Glymour (1986) that "statistics runs with a lot of philosophy, too much of it tacit, and bad philosophy is best avoided by explicitness". Yet much contemporary philosophical writing is largely turned in on itself. Disputing the validity of colleagues' positions is a fundamental mode of discourse among professional philosophers. Scientific disputes can also be sharp, but there is a greater sense that science works towards a consensus on achievable progress. Perhaps it is better to run with scientists turned philosophers, or philosophers whose inspiration derived from accurate descriptions of scientific thought and method.

Philosphers like to debate the strengths and weaknesses of 'isms' such as realism, rationalism, empiricism, or instrumentalism. An alternative form of discourse, since each such term flags important aspects of a working philosophy, is to mold these diverse but relevant aspects into a coherent overall position. Such an integrated view is necessarily complex, and hence expensive to define and communicate, thus resembling science itself. It is interesting that, although the various 'isms' of philosophy were developed through examples that overlap little with current issues in statistics, they often appear as implicit underpinnings of opposing positions in statistics. Accordingly, since complexity is necessary anyway, it may cost little to reorient statistical debates, away from one-issue confrontations, and towards more subtle difficulties.

For example, a realist position treats phenomena that science seeks to understand as having an objective hard existence in the outside world, whence science achieves objectivity by concentrating on the construction of accurate reflections of objective reality. Statisticians of this persuasion are likely to stress the desirability of explaining statistical phenomena exclusively through long run frequencies computable in principle from objective counts and measurements, with the idea of getting at real objective truth. By constrast, if realism is as much metaphor as scientific necessity, then a less fettered approach to understanding how the scientist relates to the external world may yield important insights.

A rationalist might seek to elevate correct reasoning to the position of supreme guide to correct science. One version, not much in vogue, holds that some scientific knowledge is true a priori, and is objective because we all possess it, hard-wired, from birth. Persons holding 'necessary' (a term used by Savage, 1954) views of probability conform to this type of rationalism. Other rationalists seek truth in the form of axiom systems that can be directly perceived as objectively valid and whose consequences include all the right reasoning required to generate scientific knowledge. Modern Bayesian statisticians often argue their case on such grounds. My own view is that understanding axiomatics is philosophically less important than, and certainly does not displace, efforts to explain the complex system of model construction and inference tasks that constitute reasoning about phenomena.

Empiricism enters statistics through the widespread ethic that statisticians should limit their activities to reporting and analyzing facts. A concomitant position is strongly entrenched in the field, namely, that the science of statistics is almost exclusively devoted to specifying appropriate methods of statistical analysis for each available data collection scheme. In particular, the inductive method of classical empiricism is reduced to extended study of the properties of methods. I find the absence of reasoning about actual circumstances from this widespread account of statistical practice astonishing. Another side of empiricism stresses the necessity of careful observation and experiment, and is faultless in its desirability and importance. But there is much more to the story of how science is done.

Instrumentalism asserts that the methods of science are no more than tools that make possible inferences about outcomes and quantities that are not directly ob-

served. A movement exists within statistics that stresses 'predictive inference' as the central activity, the idea being that estimation of parameters in models may be sensitive to model choice, whereas inference about not-yet-observed quantities is operationally relevant and facilitates testing of models. Again, the metaphor is appealing, but contains no hints about how to construct instruments, nor about the nature of the skills required to use them with effect.

An important philosopher whose views accord well on many dimensions with my untutored positions is Gaston Bachelard (1884–1962). Bachelard came late to philosophy and his basic writing on the philosophy of science appeared between his 1927 thesis and his 1940 appointment to a chair at the Sorbonne. His later writing emphasized psychoanalytic and aesthetic themes and was translated earlier, but his centenary year brought forth a translation of a 1934 work (Bachelard, 1984) and an excellent exposition of his scientific ideas (Tiles, 1984).

Bachelard (1984) starts by outlining the "essential complexity of the philosophy of science", so that "every man who attempts to use science makes use not of one but of two metaphysical systems. Both are natural and cogent, implicit rather than explicit, and tenacious in their persistence. And one contradicts the other." He quotes Bouty (1908): "Science is a product of the human mind, a product that conforms to both the laws of thought and the outside world. Hence it has two aspects, one subjective, the other objective; and both are equally necessary, for it is as impossible to alter the laws of the mind as it is to change the laws of the Universe."

An essential Bachelardian idea is that scientific analysis grows from a complex interaction between subject and object, between the working scientist and the world of phenomena. The nature of the interaction has evolved and will continue to evolve through history, with periods of rapid change or discontinuity. The subjectivism of science differs sharply from common sense reasoning and untrained psychological judgement. In particular, scientific reasoning makes use of modern mathematics whose constructed abstractions lead to scientific representations very distant from those of earlier times. Bachelard says that modern rationalism is non-Euclidean, symbolizing the changes that have come with modern mathematical thought. Likewise, current views of technical objectivity are non-Baconian because empirical content is not all, and progress involves more than induction from observations. Thirdly, contemporary epistemology is non-Cartesian. The social processes by which we acquire scientific knowledge do have a complex type of objectivity, but a full philosophical description must incorporate processes of active reflection, analysis, criticism, and correction that take place within and between individual subjective minds.

Bachelard does recognize causality as a fundamental driving force of modern scientific thinking, one reason being that causal thinking is able to bridge the gap between determinism and indeterminism. Causation goes deeper than deterministic explanation because statistical explanation extends the range of deterministic laws to include the explanation of mass phenomena, such as entered physics through the kinetic theory of gases and Heisenberg's uncertainty principle. Such fundamental

indeterminacies are for him nonetheless causal. Although probability is not a subject of much concern in Bachelard's writing, he touches it briefly in connection with determinism vs indeterminism, and he tentatively asserts, "The rationalizations of empirical statistics probably must proceed by establishing a correspondence between probability and frequency" (Bachelard, 1984, p. 118), thus agreeing with the non-subjectivist positions of contemporaries such as Reichenbach and von Mises. (Reichenbach (1949) goes further and asserts that such probabilism replaces causalism.)

It thus appears that Bachelard assigns probability, and its associated mathematical theory, an important role on the realist side of the rationalist/realist dualism. In the two books that I have reviewed, however, there is no sign that he gives any heed to subjective probability. It seems paradoxical that he argues forcefully for the importance of subjectivism, yet omits discussion not only of subjective probability, but of both determinate Boolean logic and nondeterminate probabilistic logic. His recognition of the power of mathematics on the realist side is not balanced by appreciation of the powerful forms of certain and uncertain reasoning that can be represented by mathematical logic and probability theory. Instead, according to Tiles (1984, p. 25) he develops "non-formal characterisations of the epistemological structure of thought and of the relation between experimental and theoretical practices, together with an account of the dynamic epistemological role of critical reflection".

Bachelard's "non-formal characterisations" are insightful and accurate, but it seems unBachelardian to leave out of the picture coexisting formal structures that are valuable in speeding deductive or computational tasks and in preventing errors to which unaided common sense reasoning is often prone. My sense is that the omission is a simple consequence of the tradition of teaching in the natural and mathematical sciences that existed in Bachelard's time, and continues dominant to this day. In fact, there is a real 'discontinuity', of the sort that Bachelard himself saw as characteristic of "progress" in fields of science, between traditionalists, whose subjectivism is strictly informal, and the small fractured minority, including myself, who argue that parts of science should be expressed in formal subjectivist terms.

It is a challenging task of historical analysis, exceeding my present knowledge and historical skills, to sort out the processes of change among leading thinkers that produced the widespread 20th century belief that subjective probability is a discredited concept. The original seminal innovators such as Jacob Bernoulli and Laplace certainly thought in subjectivist terms, and the views and methods of Laplace in particular were widely transmitted outside of mathematics, for example, to Quetelet, and to outstanding British scientists such as Maxwell and Herschel, as traced by Porter (1986). Keynes (1921, p. 172) quotes "from a letter written by Maxwell in his nineteenth year (1850)" to the effect that "the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind." Boole (1854) went far towards integrating deterministic and probabilistic logic. Soon, however, much of

the scientific community switched, in effect narrowed, from subjectivist to stochastic and frequentist conceptions of probability. The causal role of influential writers such as Cournot in France, and Ellis, Venn, and Mill in Britain, in effecting the switch deserves further analysis.

The historical puzzle does not center on probability alone, however, because neither the deterministic or probabilistic sides of Boole's paradigm took permanent hold in the working vocabularies of active members of the scientific establishments and disciplines that grew rapidly from the middle of the 19th century. Mathematical logic has flourished as an isolated specialty motivated by the desire to mathematically model mathematical proof, whereas attempts by proponents of artificial intelligence to apply the theory to external phenomena appear to have established only a few scientific toeholds. Historically, it appears that both deterministic and nondeterministic aspects of formal subjectivist reasoning retreated simultaneously into minority status, presumably for connected but as yet not well traced reasons.

My sense is that determinate and indeterminate logics should redevelop, along lines initiated by Laplace and Boole, into a canonical formal theory that will increasingly contribute to scientific progress in many fields. In particular, subjectivist probability is not an alternative to frequentist probability, but is complementary in a way that requires only a simple reinterpretation of random phenomena in mixed rationalist and realist terms, as discussed below. These issues are somewhat tangential to the main topic of causality, but understanding causality has at least as much to do with causal reasoning as with causal processes, or with rationality as with reality, hence it is important at the outset to introduce and stress the dualism of causal analysis.

3. Causality and scientific uncertainty

In many simple repeatable controlled experiments, such as those of school science courses, an action A is invariably followed by an effect B. In others, such as experiments where the observation is a particle count, or a measurement subject to error, an action A produces different possible effects B that under replication appear consistent with independent random draws from an unvarying distribution. The latter type of experiment may be said to produce statistical regularity. Thinking about such examples leads scientists to imagine that deterministic and probabilistic causal mechanisms have objective real world existence. In the probabilistic case before each experiment the outcome is uncertain, but the experimenter who possesses long run frequencies may make objective probability statements. Indeed, since no outcome is completely certain, deterministic causation can be no more than a limiting form of probabilitic causation where the probability approaches one. Thus it is easy to be seduced into thinking that the external world is fundamentally stochastic.

According to Bachelard (1984), however, "the psychology of probability remains

rather obscure, for it stands in opposition to the whole psychology of action. *Homo faber* has been hard on *homo aleator*; realism has been hard on speculation. There are physicists whose minds are closed to ideas of probability," and he notes "Lord Kelvin's strange incomprehension in this regard." Thus stochastic probability cannot rest exclusively on the realist side. The working scientist needs a working meaning for stochastic probability, relating each probability to an uncertain outcome that it aims to characterize. To me, this implies no escape from involving the subject in the formal analysis, for it is the subject who carries out the reasoning. The widespread nonrationalist tradition of teaching and thinking has encouraged many scientists to ignore the problem of interpretation, so there is much discomfort when no solution appears short of overturning tradition and associating mathematical probabilities with the subjectivist side of the basic duality, thus normalizing a formal machinery of uncertain reasoning. We then enter a world of science that is radically new for the traditionalist, even the traditionalist who has accepted that the real world is to a degree random.

In another sense the situation is made simpler, however, for the necessity of a concept of fundamental, even physical, randomness is much diminished. Specifically, there is little or no need to operate with a concept of stochastic causal mechanism, and one can retain as the basic primitive of causality the familiar image of deterministically acting causal mechanisms. Apparent randomness is explained, and so rationalized, by subjective probability, but is due to "our ignorance of the true causes" (Laplace, 1951), not to probabilistic causation *per se*. For example, at conception, the genetic makeup of a child is determined by the pairing of chromosomes from its mother and father. The unknown value of a particular marker for a newly conceived child would often be assigned a probability .5, or some different specific numerical value determined by a simple stochastic model (ignoring empirical studies that might indicate small deviations from such theoretical numbers). A traditionalist could choose to regard the gene matching process as a physical random mechanism. But this is something of a mystery, because scientific understanding of the biology is more fully reflected by a deterministic description of sperm selection and movement. When the descriptive realism is disentangled from the epistemic subjectivism, however, and neither is distorted to take over the domain of the other, the mystery disappears. The probabilities simply reflect Laplacian 'ignorance', and can be applied either before the event of conception, or after, as long as we remain ignorant. The language of chance mechanisms may convey illuminating analogies to familiar games of chance that are conventionally described as random, but one may remain agnostic and skeptical about the realism of physical randomness as a causal phenomenon.

Laplace's subjectivist attitude was widely held in the 19th century. For example, it was spelled out at length by W.F. Donkin (1851), who was Savilian Professor of Astronomy in the University of Oxford, to rebut frequentist critics. For Karl Pearson (1911), 'proof' in the establishment of 'cause and effect' is the "demonstration of overwhelming probability", and he rehearsed the Bayesian argument in detail in

Pearson (1920). I believe (Dempster, 1989) that the remarkable 20th century innovations of R.A. Fisher were fundamentally driven by the subjectivist interpretation. His was a struggle to transcend the confining Bayesian models of his 19th century forbears, but his book on inference (Fisher, 1956) can be read as a remarkable celebration of the subjectivist tradition of probabilistic thinking. Most statisticians will regard this view as wrong, even sacrilegious, but really it is commonplace when viewed merely as part of the subjectivist character of all aspects of model construction. Consider for example the obviously true proposition that Fisher continually used terms like "scientific induction" that showed his fundamental concern for both rationalist and realist aspects of the scientific enterprise.

Probabilistic uncertainty is very prominent due to its associated highly developed mathematical theory. Models that apply the theory are constructed, and often pass through stages from tentative and speculative to firm and well supported. Other uncertain aspects of scientific model construction that are even more fundamental also exhibit a progression from soft and partially formed structures to hard mathematical representations. For example, the statistician identifies and labels types of units, such as individual persons, or groups such as families or households. The statistician also identifies variables that characterize differences among units of the same type. Relations among the values of variables, whether deterministic relations holding by definition, or empirical relations with lesser degrees of precision, are likewise developed and modified over time. Knowledge structures are painfully constructed in this way and are stored in memories along with instructions for use so they can pass from person to person and generation to generation. There is no sharp distinction in the uncertain processes of scientific model construction between probabilities, that are described here as subjective, and the other equally basic aspects, such as units and variables, that traditionalists accept as completely objective. Elements of each type are constructed according to prescriptions that are both informal/formal and subjectivist/objectivist. In fact, the whole modelling enterprise is accurately perceived only as taking place against a background of interacting internal and external realms that tentatively advance and modify formal mathematical representations.

Part of the illusion that some scientific analyses are completely objective comes from contemplating circumstances that appear to approximate ideals such as perfectly controlled experiments. Actual practice, especially in statistics, always involves imperfectly controlled data collection schemes reflecting phenomena of great complexity. Preformal analysis must then proceed to sift through many hypothetical causal images that hopefully can be shaped into a network of interacting effects to explain partially the workings of apparently random phenomena. Physical science presents many good examples in the area of 'ill-posed inverse problems' (Tikhonov and Goncharsky, 1987), exhibiting a range from situations with precise deterministic mathematical models, to others with loosely formed and tentative models, in either case with random measurement errors overlaid. A good working philosophy must be prepared to cope with such a range. Statistical problem analyses are not always

satisfactory, but even speculative analyses are necessary steps on the way to more trustworthy solutions.

4. Statistical analysis and causation

A good entry point to the statistical literature on how to detect and measure causal effects is Holland (1986), including the appended discussion. As with much 20th century statistical science, the original impetus came from R.A. Fisher, in particular from his introduction at Rothamsted of basic techniques of statistical design of experiments, including such key notions as blocking and randomization (Fisher, 1935). An ongoing problem, discussed for example in Rubin's (1984) review of the writing of W.G. Cochran, is to understand how to assess limitations on the validity of inferences about causal effects, given that the conditions of ideal experimentation have not been maintained. The motivating concern is practical: to assure that nominated causal effects will stand up under replication.

Since the practical side of causal analysis does not exist in a vacuum, statisticians need also to understand and respect the preformal side. As Bachelard notes (1984, p. 112), "Scientists do not spend all their time making measurements. The seek first to understand how phenomena are interrelated and often conceptualize such interrelationships in qualitative rather than quantitative terms." Such causal thinking leads to models: "the use of causality in the sense of cause as explanation is very critical in the development of mathematical models" (Kempthorne, 1984). From models we can proceed to inference: "The inductive method exalts experiment into a position of supreme importance, but it is sometimes forgotten that the aim is generalization" (Fisher and Stock, 1915). Thus it is important to expose and clarify the connections between causal thinking and principles of statistical design, and then equally important to understand other connections, for example, to modelling and inference.

"That correlation is not causation is perhaps the first thing that must be said" (Barnard, 1982, p. 387) is a wise statement in need of clarification. A more fundamental statement is that correlation does not guarantee repeatability. To warn against confusing correlation and causation is to warn against believing that future interventions will be effective. But similar caution is in order for predictions generally. An effect visible in immediate data need not replicate in new data, whether the effect refers to noncausal predictions or to predictions of causal effects of interventions. In an important sense, therefore, the statistical principles that explain why it is difficult to make real world inductions from empirical correlation are not closely tied to the concept of causation. I believe this explains the paradoxical insistence of Holland (1986, p. 959) that the proper concern of statisticians is with "studying the effects of causes rather than the traditional approach of trying to define what the cause of a given effect is." We need also to be concerned with the latter, but through principles complementing those ably reviewed by Holland.

Induction from an empirical relation, such as from a scatterplot of (x, y) pairs,

assumes knowledge of another sort, namely, of a relation between the set of units that define the data and the target unit or set of units to which the inference will be made. Loosely, each set is assumed to be representative of a common population. An ideal form of representativeness is the assumption of random sampling, a subjectivist probability assessment asserting that the observed units were *a priori* exchangeable with any subset of a specified population of units. A canonical form of induction is a Bayesian posterior distribution for the y value of a target unit, given the x value for the unit and given the (x, y) pairs for the data set. If the target is a set of units, the inferences may be directed at individual members, or at aggregates over the target sample, or in the limit at a full target population. If inferences are drawn about a single target unit, it is critical that the unit is judged to be a random draw from the same population represented by the data set, or at least a random draw from the subset of the population conditioned on the x value of the target unit. Throughout, formal probabilistic inferences require precise inputs of formal probabilistic knowledge (or 'assumptions' in traditional terminology).

In an experiment to assess causal effects, there are at least two data sets, where in one the units receive treatment t ('the treatment') and in the other receive treatment c ('the control'). Each set must be representative of an available population in a precise sense that permits formal induction. The underlying reason for R.A. Fisher's advocacy of randomized treatment assignment is precisely that it affords the subjective probability assessor the luxury of a real world basis plainly visible to fellow scientists. It thus becomes possible to make uncertain inferences about the consequences that would have ensued had c been applied to the set that actually received t, and vice versa. These inferences, which were criticized by Glymour (1986) as counterfactual, are not however the main point. A greater practical issue concerns prediction of the consequences of applying treatments t or c in the future to a further set of representative target units, whence the desirability, often difficult to achieve, of random sampling of experimental units. For example, the validity of policy analyses that inform institutional decisions rests on the ability to produce appropriately predictive inferences.

Comparative predictions can be assessed with varying degrees of specificity. The most specific treatment comparison is a prediction for a single prospective target unit with known x, but comparisons for different levels of aggregation and varying degrees of knowledge of x can be interesting and important. A key observation is that causal thinking points first at single units: while a causal action may be broad in the sense of affecting many units simultaneously, and while a causal story sometimes nees the complexity of allowing for interactions among units, the basic mental picture is of a reactive process that occurs within a single unit. Hence the term causal effect refers in a direct sense to single units, and only by aggregation to sets of units. Statisticians by contrast are conditioned to think first about description, and description starts from population aggregates. On this point, I believe that the beautifully clear Rubin-Holland account of causal inference, as concisely presented in Rubin (1986) and the rejoinder of Holland (1986), may seem to promise more than it can

deliver. The robustness claimed by Rubin (1978) as a benefit of randomization for a Bayesian is most effective at the level of comparing aggregates, but the later extension to causal effects at the level of individual units requires the strong stable-unit-treatment-value-assumption (SUTVA) of Rubin (1980). The theory is clear, but the difficulty of providing an objective basis for this assumption contrasts sharply with the objective basis of the exchangeability assumption provided by randomization.

The simplicity of SUTVA also stands in contrast with the spirit behind the dictum of Fisher quoted by Cochran (1965, p. 252): "Make your theories elaborate." As Cochran remarks of studies of smoking, "In the largest studies, we can compare the death rate (i) of men who smoked different amounts for the same time, (ii) among smokers of the same amount, of men who had beens moking for different lengths of time, (iii) of exsmokers and current smokers of the same amount, (iv) among ex-smokers, of those who had previously smoke different amounts, and (v) among ex-smokers of the same amount, of those who had stopped recently and those who had stopped for longer periods." Cochran remarks that predictions of causal effects across such varied categories of units can be of great value in sorting through hypotheses about alternative causal mechanisms. But the complexity of the desired inferences greatly increases the modelling task of the Bayesian statistician. Scientific inference from comparative studies is difficult precisely because an objective basis for such models can be constructed only with much patient labor.

An important truth emphasized by Rubin (1978) is that randomization, when achievable, assures prior independence of treatment assignment and other attributes of the units, including attributes that reflect causal factors other than the experimental treatment. Thus randomization is one guarantor of the validity of mathematical representations of interacting effects of several observed causal factors, and similarly allows modelling of effects of unobserved causal factors as random error, both in a manner that can often extend to target units to be treated in the future. There is a catch, however, because the posited statistical modelling requires not only good statistical control of unit selection, treatment, and observation processes, but also demands samples sufficiently large that the modelling tasks can be carried out with acceptable levels of objectivity. When the phenomena are complex and dynamic, the technical problems of achieving and analyzing adequate samples are often beyond the current state of the methodologies involved.

Statistical modelling of complex systems has advanced rapidly in recent decades. Important current work is by Wermuth and Lauritzen (1989) on modelling, and Lauritzen and Spiegelhalter (1988) on computation. The current state of dynamic statistical modelling is surveyed in Spall (1988). Deterministic and stochastic models that postulate separate observables and unobservable 'state' vectors have long been used in physical and engineering sciences (Tikhonov and Goncharsky, 1987) and should form an important growth area for statistical modelling of biological and behavioral phenomena. The type of subjectivist/objectivist modelling advocated in this paper has a natural fit with the type of causal thinking that was regarded as basic by leading practitioners such as Cochran and Fisher.

5. Examples

The picture of statistical problem analysis developed in this paper is distant from the picture created by most of the literature of statistics with its emphasis on the selection and application of methods and the study of properties of methods. Instead, I have been stressing mathematical modelling of reality, and rationalist interpretations of models, so that the introduction of causal thinking becomes natural. I close with some examples where causality is primary, and where the advocated principles are natural, and probably necessary, tools of analysis.

5.1. Breast cancer and alcohol

The abstract of Willett et al. (1987) reads, "In 1980, 89 538 US women 34 to 59 years of age, with no history of cancer, completed a dietary questionnaire that included the use of beer, wine, and liquor. During the ensuing four years, 601 cases of breast cancer were diagnosed among cohort members. Among the women consuming 5 to 14 g of alcohol dialy (about 3 to 9 drinks per week), the age-adjusted relative risk of breast cancer was 1.3 (95 percent confidence limits, 1.1 and 1.7). Consumption of 15 g or more per day was associated with a relative risk of 1.6 (95 percent confidence limits, 1.3 and 2.0; Mantel extension chi for linear trend = 4.2; $P<0.0001$). Adjustment for known breast cancer risk factors and a variety of nutritional variables did not materially alter this relation. Significant associations were observed for beer and liquor when considered separately. Among women who were without risk factors for breast cancer who where under 55 years of age, the relative risk associated with consumption of 15 g of alcohol a day was 2.5 (95 percent limits, 1.5 and 4.2). These prospective data derived from measurements of alcohol intake recorded before the diagnosis of breast cancer confirm the findings of several previous case-control studies. Viewed collectively, they suggest that alcohol intake may contribute to the risk of breast cancer."

It appears that the authors have taken a complex view, looking for alternative explanations of their findings, establishing consistency across subgroups, and indicating a suggestion of a dose-response relation that would support a biological mechanism. Yet Feinstein (1988) strongly criticizes Willett et al. for the lack of "fundamental scientific standards used to specify hypotheses and groups, get high-quality data, analyze attributable actions, and avoid detection bias." Feinstein indicates also that the message from competing epidemiological investigations is more ambiguous than the abstract admits, or even then the body of the paper allows. The dispute is ongoing, and is unlikely to be resolved in a way that fully vindicates either side. Further data, better analyses, and improved biological understanding will change the terms of the debate.

What are the lessons for statisticians? The absence of randomized treatment assignment is a necessity, and the sample is not a random sample but is probably adequately representative of middle class cohorts of urban US women. Surely

medical research is better off with the data than without, even when weighed in the cost scale against competing priorities. Perhaps modest improvements can be made towards Feinstein's 'standards', and certainly subjectivist assessments and related computations could serve to quantify the possible effects of alleged defects.

At a more fundamental level, there are substantial opportunities for developing analyses that go much further towards meeting the underlying science. Attention is being given, in this and other areas of epidemiology, to the problems of measuring exposure to potential causal agents. The problem is not simply the accuracy of the diet questionnaire in reflecting alcohol consumption in a specific year, but includes in principle the task of assessing true alcohol consumption as a function of time back to puberty for each study member. It is equally important to tie analyses to models of carcinogenesis as a multistage process. As reviewed by Armitage (1985), several approaches to modelling are under active development. In particular, the type of model advocated by Moolgavkar (1986) has been related to epidemiological incidence data and to biological phenomena at the genetic level. In principle, there is no bar to developing Bayesian analyses of the data set of Willett et al. that treat the underlying phenomenon as generated by Moolgavkar's model and experiment with subjectivist assessments of model parameters and true exposure histories. Developing feasible computational procedures might take several man years of highly skilled labor, and the immediate payoff regarding alcohol and breast cancer might not be large since the analyses of Willett et al. presumably do exhibit the main statistical relations in the data. The long term benefits are more promising, since the use of subjectivist modelling permits the study of detailed causal mechanisms of initiation and promotion that may suggest improved hypotheses and empirical studies to test them.

5.2. *Employment discrimination and statistical science*

For several decades, the US Government has sought through legislation and regulation to equalize employment opportunities for women and minorities as compared to white males. In particular, sanctions can be imposed through judicial proceedings against corporations that are found to engage in discriminatory practices in hiring, advancement, or remuneration. Such practices can be identified at the level of individual employment decisions, but plaintiffs typically attempt to use statistics to demonstrate a consistent pattern of discrimination across large segments of a corporation. The positions taken on interpreting the data reflect the scientific standards offered by competing expert witnesses, legal arguments and reasoning put forth by opposing lawyers, and partisan interpretations from plaintiffs and defendants often reflecting political attitudes and paying little heed to reason.

The outlook for good statistical science in this area is bleak, even assuming the existence of statistical standards resistant to legal and political manipulation. The sample generally has no scientific credentials, typically being a set of employer administrative records that happen to be available. The treatment whose causal effect

is at issue is often a largely unrecorded process, including nonscientific subjective judgements by a decision-maker, leading up to an outcome which is a recorded employment decision. In the epidemiology example a direct measurement of treatment dose is available, whereas in discrimination studies gender and minority status themselves are used as proxies for treatment assignment. Because the goal is to show a stable association between the dose of discriminatory behavior and the proxy, it is clear that any analysis will be heavily model dependent. Nor are simple hypotheses like SUTVA *a priori* plausible.

Dempster (1984) pointed out some of the difficulties, much as Feinstein did for the epidemiological example. In Dempster (1988), I tried to set out the elements of a scientific Bayesian analysis in a simple situation. It should be evident that courts are interpreting data in favor of plaintiffs beyond permissible limits of scientific credibility.

5.3. *Retrospective analysis of a possible adverse drug reaction*

Lane (1989) illustrates a subjectivist Bayesian methodology using the case of a French woman age 38 who had lived in Gabon for about 2 years and who developed hepatitis and died in March, 1984, about 2 months after her first symptoms were noted. About 3 months before her death, she had switched drugs for malaria prophylaxis from chloroquine to amodiaquine. Lane developed his analysis as moderator of a panel of experts from the Montreal medical research community. There were two main hypotheses: (i) that the hepatitis was a direct adverse reaction to amodiaquine, this being the reason for the original study of the case by experts in France where the patient was taken 5 days before her death, and (ii) that the hepatitis was a form of viral hepatitis labelled NAND that is fairly common in Gabon and not ruled out by the tests recorded for the patient. The panel assessed relative odds of 0.24 for (i) vs (ii) given the bare facts without details of the clinical history for the three months course after initiation of amodiaquine, and a further likelihood odds multiplier, given the details, of 6.3, yielding final odds of 1.5 in favor of (i). The clinical details included a record of the advance, decline, and fatal advance of the disease in relation to the timing of an interruption of amodiaquine treatment, and also tests relevant to hepatitis. Lane gives many fascinating details about the wide range of medical knowledge invoked and how it fits into the various stages of his processes of problem formulation and analysis.

The presentation is limited to one question: which of two nominated causes actually operated? Other retrospective causality assessments, for example concerning the Challenger space shuttle disaster or the Chernobyl nuclear plant explosion, start with a broad range of possibilities that quickly move to a single main candidate as data-gathering and analysis proceed. We are thus reminded that the informal subjective phase of problem formulation often stops before any question of formal probability assessment emerges. In the Challenger case, there was a convincing demonstration after the fact that a careful probability risk assessment ought to have

been done before launch (Dalal et al., 1989). Likewise, when competing hypotheses survive careful informal analysis, the explicitness of formal analysis is likely to result in fewer errors of omission and of logic than thoroughgoing informal subjectivism.

The explicitness of Lane does not guarantee that other analysts would not have been led to to consider and use other epidemiological and theoretical knowledge than was used by the Montreal panel. Lane does discuss two replications of his methodology in France applied to the same case without major differences. Possibly, however, some quite different vantage point would have led to a different analysis. Suppose the intial focus had been on assessing all the evidence about hepatitis as a side effect of amodiaquine. This seems a more natural place for a statistician to start, and might well have raised issues that were not raised by the medical panels, that could in turn have changed the subsequent course of analysis. Thus nonmedical mathematical scientists might have useful roles beyond that illustrated by Lane. For now, the conclusion on the particular patient is close to a toss-up, until such time as new knowledge and arguments significantly shift the analysis.

5.4. The greenhouse and climate change

As we progress through the examples 5.1 to 5.2 to 5.3, the concept of a relevant statistical sample becmes increasingly indefinite, whence the possibility of valid inferences directly from statistical analysis becomes more remote. In the case of possible effects of the buildup of atmospheric CO_2 and other greenhouse gases, there is by definition no possibility of replicated systems. Only the actual system is available, and we must predict alternative future climates given alternative possible futures of atmospheric buildup. Nevertheless, the amount of available data from the past is huge and growing, as is the store of reliable knowledge about physical processes that affect climate. Thus the situation is reversed from the ideal of the Rubin-Holland analysis, where randomization alone carries most of the burden of assessing causal effects. Here, by contrast, the desired inferences about the effects of continued buildup of greenhouse gases is heavily dependent on modelling and analysis, needing strong inputs of prior knowledge, of a single highly structured and multivariate time series.

Just as induction from a sample of units to further sample units depends on a representativeness assumption, most often a random sampling hypothesis, so does prediction from an observed time series depend on a time invariance assumption, namely, that any stochastic mechanism postulated to underly the observations should be invariant under time shifts. The anthropogenic inputs of greenhouse gases over the past century or so have been accelerating in a way that has not been replicated in available records, and hence, for example, a bivariate time series model of global temperature and atmospheric CO_2 would need to be treated as a very questionable guide to forecasting. The problem does not lie with time series methods *per se*, but rather with their premature application before the time homogeneity assumption is

plausible. If there is a reason for optimism that prediction is possible, it must come from a belief that the basic principles governing the biosphere are time homogeneous. Consequently, modelling must be carried to levels of detail sufficiently fine that the underlying time series can be trusted for reliable prediction because they are based on accumulations of tested scientific knowledge and understanding. An excellent introduction and review of current knowledge and modelling efforts on the greenhouse effect is given by Kondratyev (1988). Perhaps the models are not yet sufficiently advanced that acceptable forecasts are possible, but it is not too soon for statisticians to be studying the models in order to understand how valid time series models and associated Bayesian forecasts could develop from the underlying science.

References

Armitage, P. (1985). Multistage models of carcinogenesis. *Environmental Health Perspectives* **63**, 195–201.

Bachelard, Gaston (1984). *The New Scientific Spirit*. Beacon Press, Boston, MA (Originally published as *Le Nouvel Esprit Scientifique*, Alcan, Paris, 1934).

Barnard, G.A. (1982). Causation. In: S. Kotz, N. Johnson and C. Read, Eds., *Encyclopedia of Statistical Sciences* Vol 1. 387–389.

Boole, George (1854). *The Laws of Thought*. Reprinted by Dover Publications, New York, 1958.

Bouty, Edmond (1908). *La Verité Scientifique*.

Cochran, W.G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. Ser. A* **128**, 234–265.

Dalal, S.R., E.B. Fowlkes and Bruce Hoadley (1989). Risk analysis of the space shuttle: The pre-*Challenger* prediction of failure. *J. Amer. Statist. Assoc.* **84**, 945–957.

Dempster, A.P. (1984). Alternative models for inferring employment discrimination from statistical data. In: P.S.R.S. Rao and J. Sedransk, Eds., *W.G. Cochran's Impact on Statistics*. Wiley, New York.

Dempster, A.P. (1988). Employment discrimination and statistical science. *Statist. Sci.* **3**, 149–161.

Dempster, A.P. (1989). Bayes, Fisher, and belief functions. In: S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, Eds., *Bayesian Likelihood Methods in Statistics and Econometrics*. North-Holland, Amsterdam.

Donkin, W.F. (1851). On certain questions relating to the theory of probability. *Philosophical Magazine [Fourth Series]* **1**, 353–368; **1**, 458–466; **2**, 55–60.

Feinstein, Alvan R. (1988). Scientific standards in epidemiologic studies of the menace of daily life. *Science* **242**, 1257–1263.

Fisher, R.A. (1918). The causes of human variability. *Eug. Rev.* **10**, 46–61.

Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

Fisher, R.A. and C.S. Stock (1915). Cuénot on preadaption: a criticism. *Eug. Rev.* **7**, 46–61.

Glymour, Clark (1986). Comment: statistics and metaphysics. *J. Amer. Statist. Assoc.* **81**, 964–966.

Good, I.J. (1961/62). A causal calculus. *Brit. J. Philos. Sci.* **11**, 305–318; **12**, 43–51; **13**, 88 (reprinted in *Good Thinking*, University of Minnesota Press, 1983).

Good, I.J. (1988). The interface between statistics and the philosophy of science. *Statist. Sci.* **3**, 386–397.

Herschel, John (1850). Quetelet on probabilities. *Edinburgh Rev.* **92**, 1–57.

Holland, Paul W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81**, 945–960.

Kempthorne, Oscar (1984). Statistical methods and science. In: P.S.R.S. Rao and J. Sedransk, Eds., *W.G. Cochran's Impact on Statistics*. Wiley, New York.

Keynes, John Maynard (1921). *A Treatise on Probability*. Macmillan, London (reprinted Harper and Row, New York, 1962).

Kondratyev, K.Ya. (1988). *Climate Shocks: Natural and Anthropogenic*. Wiley, New York.

Lane, David A. (1989). Subjective probability and causality assessment. *Applied Stochastic Models and Data Analysis* **5**.

Laplace, P.S. (1951). *A Philosophical Essay on Probabilities*. Dover Publications, New York (translated from the 6th French edition, Courcier, Paris, 1840).

Lauritzen, S.L. and D.J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. Ser. B* **50**, 157–194.

Moolgavkar, Suresh H. (1986). Carcinogenesis modeling: from molecular biology to epidemiology. *Ann. Rev. Public Health* **7**, 151–169.

Pearson, Karl (1911). *The Grammar of Science* (3rd ed.). Walter Scott, London (reprinted Peter Smith, Gloucester, MA, 1969).

Pearson, Karl (1920). The fundamental problem of practical statistics. *Biometrika* **13**, 1–6.

Porter, Theodore M. (1986). *The Rise of Statistical Thinking 1820–1900*. Princeton University Press, Princeton, NJ.

Reichenbach, Hans (1949). *The Theory of Probability. An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability* (2nd ed.), University of California Press, Berkeley, CA.

Rubin, D.B. (1978). Bayesian inference for random effects: the role of randomization. *Ann. Statist.* **6**, 34–58.

Rubin, D.B. (1980). Comment. *J. Amer. Statist. Assoc.* **75**, 591–593.

Rubin, D.B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In: P.S.R.S. Rao and J. Sedransk, Eds., *William G. Cochran's Impact on Statistics*. Wiley, New York.

Rubin, D.B. (1986). Comment. Which ifs have causal answers? *J. Amer. Statist. Assoc.* **81**, 961–962.

Russell, Bertrand (1948). *Human Knowledge. Its Scope and Limits*. Simon and Schuster, New York.

Savage, L.J. (1954). *The Foundations of Inference*. Wiley, New York.

Shafer, Glenn (1986). Savage revisited. *Statistical Science* **1**, 463–485.

Spall, James C. (Ed.). (1988) *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker, New York.

Suppes, Patrick (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Suppes, Patrick (1988). Comment: causality, complexity and determinism. *Statist. Sci.* **3**, 398–400.

Tikhonov, A.N. and A.V. Goncharsky (Eds.). (1987). *Ill-Posed Problems in the Natural Sciences*. Mir, Moscow.

Tiles, Mary (1984). *Bachelard: Science and Objectivity*. Cambridge University Press, Cambridge.

Wermuth, Nanny and Steffen L. Lauritzen (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser. B* **52**, 21–50.

Willett, W.C., M.J. Stampfer, G.A. Colditz, B.A. Rosner, C.H. Hennekens and F.E. Speizer (1987). Moderate alcohol consumption and the risk of breast cancer. *New England J. of Medicine* **316**, 1174–1180.