In *Journal of American Statistical Association* (with discussions by G. Casella, D. Cox, S. Greenland, J. Pearl, J. Robins, D. Rubin, S. Schwartz, G. Shafer, and L. Wasserman), Vol. 95, No. 450, 428–435, June 2000.

TECHNICAL REPORT
R-269
April 2000

The Logic of Counterfactuals in Causal Inference

(Discussion of 'Causal Inference without Counterfactuals' by A.P. Dawid)

Judea Pearl
Cognitive Systems Laboratory
Departments of Computer Science and Statistics
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

Background

The field of statistics has seen many well-meaning crusades against threats from metaphysics and other heresy. In its founding prospectus of 1834, the Statistical Society of London has resolved "... to exclude carefully all Opinions from its transactions and publications—to confine its attention rigorously to facts." This clause was officially struck out in 1858, when it became obvious that facts void of theory could not take statistics very far [Annals, 1934, p. 16]

Karl Pearson launched his own metaphysics "red-scare" about causality in 1911: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect" [Pearson, 1911, p. *iv*]. Pearson's objection to theoretical concepts such as "matter" and "force" was so fierce and his rejection of determinism so absolute that he consigned statistics to almost a century of neglect within the study of causal inference. Philip Dawid was one of a handful of statisticians who boldly protested the stalemate over causality: "Causal inference is one of the most important, most subtle, and most neglected of all the problems of statistics" [Dawid, 1979].

In the past two decades, owing largely to progress in counterfactual, graphical, and structural analyses,[1] causality has been transformed into a mathematical theory with well-defined semantics and well-founded logic, and many practical problems that long were regarded as either metaphysical or unmanageable can now be solved using elementary mathematics. In the paper before us, Professor Dawid welcomes the new progress in causal analysis but expresses mistrust of the quasi-deterministic methods by which this progress has been achieved.

Attitudes of suspicion toward counterfactuals and structural equation models are currently pervasive among statisticians, and Professor Dawid should be commended for bringing such concerns into the open. By helping to dispel misconceptions about counterfactuals

[1] See Pearl (2000) for a gentle introduction to the counterfactual, graphical, and structural-equation approaches to causality.

Dawid's paper may well have rescued statistics from another century of stagnation over causality.

The empirical content of counterfactuals

The word "counterfactual" is a misnomer. Counterfactuals carry as clear an empirical message as any scientific laws, and indeed are fundamental to them. The essence of any scientific law lies in the claim that certain relationships among observable variables remain invariant when the values of those variables change relative to our immediate observations. For example, Ohm's law ($V = IR$) asserts that the ratio between the current (I) and the voltage (V) across a resistor remains constant for all values of I, including yet unobserved values of I. We usually express this claim in a function, or hypothetical sentence:[2] "had the current in the resistor been I (instead of the observed value I_0) the voltage would have been $V = I\frac{V_0}{I_0}$," knowing perfectly well that there is no way to simultaneously measure I and I_0. Such sentences appear to be *counterfactual*, because they deal with unobserved quantities that differ from (hence seem to contradict) those actually observed. Nonetheless, this circumstantial nonobservability and apparent contradiction do not diminish whatsoever our ability to submit physical laws to empirical test. Scientific methods thrive on attempts to confirm or falsify the predictions of such laws.

The same applies to stochastic processes (or data-generation models), usually written in the form of functional relations $y = f(x, u)$, where X and U stand for two sets of random variables, with joint distribution $P(x, u)$, and f is a function (usually of unknown form) that determines the value of the outcome $Y = y$ in terms of observed and unobserved quantities, $X = x$ and $U = u$. To see how counterfactuals and joint probabilities of counterfactuals emerge from such a stochastic model, let us consider a simple case where Y and X are binary variables (e.g., treatment and response) and U an arbitrary complex set of all other variables that may influence Y. For any given condition $U = u$, the relationship between X and Y must be one of the (only) four binary functions:

$$\begin{array}{llll} f_0 : y = 0 & \text{or } \{Y_0 = 0, Y_1 = 0\} & f_2 : y \neq x & \text{or } \{Y_0 = 1, Y_1 = 0\} \\ f_1 : y = x & \text{or } \{Y_0 = 0, Y_1 = 1\} & f_3 : y = 1 & \text{or } \{Y_0 = 1, Y_1 = 1\} \end{array} \tag{1}$$

As u varies along its domain, the only effect it can have on our model is to switch the relationship between X and Y among these four functions. This partitions the domain of U into four equivalence classes, where each class contains those points u that correspond to the same function. The probability $P(u)$ thus induces a probability function over the potential-response pairs $\{Y_0, Y_1\}$ shown in Eq. (1). This construction is the inverse of the one discussed in (Dawid's) section 13; we start with genuine concomitants U, and they turn into jointly distributed counterfactual concomitants $\{Y_0, Y_1\}$ that Dawid calls metaphysical and fatalistic.

Admittedly, when u stands as the identity of a person, the mapping of u into the pair $\{Y_0, Y_1\}$ appears horridly fatalistic, as if that person is somehow doomed to react in a predetermined way to treatment ($X = 1$) and no-treatment ($X = 0$). However, if we view u

[2] Every mathematical function is interpreted hypothetically, and the study of counterfactuals is merely a study of standard mathematical functions.

as the sum total of all experimental conditions that might possibly affect that individual's reaction, including biological, psychological and spiritual factors, operating both before and after the application of the treatment, then the mapping is seen to evolve reasonably and naturally from the functional model $y = f(x, u)$. This quasi-deterministic functional model mirrors Laplace's conception of nature [Laplace, 1814], according to which nature's laws are deterministic, and randomness surfaces merely due to our ignorance of the underlying boundary conditions. (The structural equation models used in economics, biology and stochastic control are typical examples of Laplacian models.) Dawid detests this conception. This is not because it ever failed to match macroscopic empirical data (only quantum-mechanical phenomena exhibit associations that might conflict with the Laplacian model), but because it appears to stand contrary to "our familiar statistical framework and machinery" (Section 7). I fail to see why a framework and machinery that did not exactly excel in the causal arena should be deprived of enhancement and retooling.

Empiricism versus identifiability

Dawid's empiricism is summarized in the abstract of his paper:

> "By definition, we can never observe such [counterfactual] quantities, nor can we assess empirically the validity of any modeling assumption we may make about them, even though our conclusions may be sensitive to these assumptions."

This warning isn't entirely accurate. Many counterfactual modeling assumptions do have testable implications: for example, exogeneity (or ignorability) $(Y_x \perp\!\!\!\perp X)$ and monotonicity $(Y_t(u) \geq Y_c(u))$ can each be falsified by comparing experimental and nonexperimental data [Pearl, 2000, Chapter 9]. More importantly, the warning is either empty or self-contradictory. If our conclusions have no practical consequences, then their sensitivity to invalid assumptions is totally harmless, and Dawid's warning is empty. If, on the other hand, our conclusions do have practical consequences, then their sensitivity to assumptions automatically makes those assumptions testable, and Dawid's warning turns contradictory.

The two queries about aspirin and headache, which Dawid uses to distinguish effects-of-causes from causes-of-effects ("sheep" from "goats"), may serve well to illustrate the inconsistency in Dawid's philosophy. The two queries are:

I. I have a headache. Will it help if I take aspirin?

II. My headache has gone. Is it because I took aspirin?

Letting $X = 1$ stand for "taking aspirin" and and $Y = 1$ stand for "having headache" (after half hour, let us say), the counterfactual expressions for the probabilities of these two queries read:

$$\begin{aligned} Q_I &= P(Y_1 = 0) - P(Y_0 = 0) \\ Q_{II} &= P(Y_0 = 1 | X = 1, Y = 0) \end{aligned} \tag{2}$$

In words, Q_{II} stands for the probability that my headache would have stayed had I not taken aspirin $(Y_0 = 1)$, given that I did, in fact, take aspirin $(X = 1)$ and the headache has gone ($Y = 0$). (We restrict the population to persons who have headaches prior to considering aspirin). Dawid is correct in stating that the two queries are of different types, and the language of counterfactuals displays this difference and its ramifications in vivid mathematical form. By examining their respective formulas, one can immediately detect that Q_{II} is conditioned on the outcome $Y = 0$, whereas Q_I is unconditioned. This implies that some knowledge of the functional relationship (between X and Y) must be invoked in estimating Q_{II} [Balke and Pearl, 1994]. I challenge Dawid to express Q_{II}, let alone formulate conditions for its estimation in a counterfactual-free language.[3] However, what is puzzling in Dawid's paper is that he considers Q_{II} to be, on one hand, valid and important (Section 3) and, on the other hand, untestable (Section 11); the two are irreconcilable. If Q_{II} is valid and important, then we should expect the magnitude of Q_{II} to affect some future decisions, and we can then use the correctness of those decisions as a test (hence, interpretation) of the empirical claims made by Q_{II}. What are those claims and how can we test them?

According to the interpretation given in the previous section, counterfactual claims are merely conversational shorthand for scientific predictions. Hence, Q_{II} stands for the probability that a person will benefit from taking aspirin in the *next* headache episode, given that aspirin proved effective for that person in the past (i.e., $X = 1, Y = 0$). Therefore, Q_{II} is testable in sequential experiments where subjects reaction to aspirin is monitored repeatedly over time. (We need to assume that a person's characteristics do not change over time, an assumption that is testable in principle.) In such tests we can easily verify whether subjects who have had one positive experience with aspirin ($X = 1, Y = 0$) have a higher than average probability of benefiting from aspirin in the future.

I have argued elsewhere [Pearl, 2000, p. 217] that counterfactual queries of type II are the norm in practical decision making, whereas causal effect queries (type I) are the exception. The reason is that decision-related queries are usually brought into focus by observations that could be modified by the decision (e.g., a patient suffering from a set of symptoms). The case-specific information provided by those observations is essential for properly assessing the effect of the decision, and conditioning on these observations leads to queries of type II, as in Q_{II}. The Bayesian approach proposed by Dawid cannot properly handle conditioning on factors that are affected by the treatment,[4] and thus deprives us of answering the most common type of decision-related queries.

I agree with Dawid that certain assumptions needed for identifying causal quantities are not easily understood (let alone ascertained) when phrased in counterfactual terms. Typical examples are assumptions of ignorability [Rosenbaum and Rubin, 1983], which involve conditional independencies among counterfactual variables. However, this cognitive difficulty comes not because counterfactuals are untestable but because dependencies among counterfactuals are derived quantities that are a few steps removed from the way we conceptualize cause-effect relationships. To overcome this difficulty, a hybrid form of analysis can be used,

[3] For background information, the identification of Q_I requires exogeneity (i.e., randomized treatment), whereas that of Q_{II} requires both exogeneity and monotonicity; both assumptions have testable implications [Pearl, 2000, p. 294]. Epidemiologists are well aware of the difference between Q_I and Q_{II} (they usually write $Q_{II} = Q_I/P(Y = 0|X = 1)$), though the corresponding identification conditions for Q_{II} are often not spelled out as clearly as they could [Greenland and Robins, 1988].

[4] Detailed dynamic models or temporally indexed data for every conceivable set of observations would be needed for specifying the probabilities in the decision trees of such analysis.

in which assumptions are expressed in the friendly form of functional relationships (or diagrams), and causal queries (e.g., Q_{II}) are posed and evaluated in counterfactual vocabulary [Pearl, 2000, p. 215-7, 231-4]. Functional models, in the form of nonparametric structural equations, thus provide both the formal semantics and conceptual basis for a complete mathematical theory of counterfactuals.

In Section 5.4, Dawid restates his empiricist philosophy in the form of a requirement which he calls *Jeffreys's Law*:

> "...mathematically distinct models that cannot be distinguished on the basis of empirical observation should lead to indistinguishable inference."

This requirement reads like a tautology; If two models entail two distinguishable inferences, and if the difference between the two inferences matters at all, then the two models can easily be distinguished by whatever (empirical) criterion we use to distinguish the two inferences. Dawid may have meant the following:

> "...mathematically distinct models that cannot be distinguished on the basis of past empirical observation should lead to indistinguishable inference regarding future observation (which may be obtained under new experimental conditions)."

This is none other but the requirement of identifiability (see e.g., [Pearl, 1995]). It requires, for example, that if our data are nonexperimental, then two models that are indistinguishable on the basis of those data entail the same value of the average causal effect (ACE) – a quantity that is discernible in experimental studies. It likewise requires that, if our data come from static experiments, then two models that are indistinguishable on the basis of those data entail the same value of Q_{II} – a quantity that is discernible in sequential experiments.

If the aim of Dawid's empiricism is to safeguard identifiability, his proposal would be welcome by all causal analysts, including adventurous counterfactualists. Unfortunately, careful reading of his paper shows that David aims at imposing an overly restrictive and unworkable type of safeguards, a type that has been outmoded in almost every branch of science.

Pragmatic versus dogmatic empiricism

The requirement of identifiability, as just stated, is a restriction on the type of queries we may ask (or inferences we may make) and not on the type of models we may use. And this brings us to the difference between pragmatic and dogmatic empiricism. A pragmatic empiricist insists on asking empirically testable queries, but leaves the choice of theories to convenience and imagination; the dogmatic empiricist insists on positing only theories that are expressible in empirically testable vocabulary. As an extreme example, a strictly dogmatic empiricist would shun the use of negative numbers, because negative quantities are not observable in isolation. For a less extreme example, a pragmatic empiricist would welcome the counterfactual model of individual causal effects (ICE) (see Section 5.2) as long as it leads to valid and empirically testable estimation of the quantity of interest (e.g., ACE). Dawid rejects this model a-priori because it starts with unobservable unit-based counterfactual terms, $Y_t(u)$ and $Y_c(u)$, and thus fails the dogmatic requirement that the entire analysis, including all auxiliary symbols and all intermediate steps, "involve only terms subject to empirical scrutiny". What we gain by this prohibition, according to Dawid, is protection from asking nonidentifiable queries. His proposal, in the form of Bayesian decision theory, indeed ensures that we do not ask certain forbidden questions, but unfortunately, it also ensures that we never ask or answer important questions (such as Q_{II}) that cannot be expressed in his restricted language. It is a stifling insurance policy, analogous to banning division from arithmetics in order to protect us from dividing by zero.[5]

Science rejected this kind of insurance long ago. The Babylonians astronomers were masters of black-box prediction, far surpassing their Greek rivals in accuracy and consistency [Toulmin, 1961, pp. 27–30]. Yet Science favored the creative-speculative strategy of the Greek astronomers which was wild with metaphysical imagery: circular tubes full of fire, small holes through which the fire was visible as stars, and hemispherical earth riding on turtle backs. It was this wild modeling strategy, not Babylonian rigidity, that jolted Eratosthenes (276-194 BC) to perform one of the most creative experiments in the ancient world and measure the radius of the earth.

This creative speculate-test-reject strategy (which is my understanding of Popperian empiricism) is practiced throughout science because it aims at understanding the mechanisms behind the observations, and thus gives rise to new questions and new experiments, which eventually yield predictions under novel sets of conditions. Quantum mechanics was invented precisely because J.J. Thomson and others took deterministic classical mechanics very seriously, and boldly asked "metaphysical" questions about physical properties of electrons when electrons were unobservable. The language of counterfactuals, likewise, enables the statistician to pose and reject a much richer set of 'what if' questions than does the language of Bayesian decision theory. Giving up this richness is the price we would pay for Dawid's insurance.

Counterfactuals as instruments

Dawid reports (end of Section 10.2) that the bounds for causal effects in clinical trials with imperfect compliance [Balke and Pearl, 1997] are "sheep-like", namely, valid, meaningful and safe even for counterfactually-averse statisticians. Ironically, when we examine the conditional probabilities that achieve those bounds, we find that they represent subjects with deterministic behavior, *compliers, never-takers* and *defiers*, precisely the kind of behavior that Dawid rejects as "fatalistic" (Section 7.1). The lesson is illuminating: even starting with the best sheep-like intentions, there is no escape from counterfactuals and goat-like determinism in causal analysis.

This lesson leads to a new way of legitimizing counterfactual analysis in the conservative circles of statistics. Researchers who mistrust the quasi-deterministic models of Laplace (i.e., $y = f(x, u)$) can now view these models as limit points of a space of nondeterministic models $P(y|x)$ constrained to agree with the observed data. Accordingly, the mistrustful analysis of counterfactuals can now be viewed as a benign analysis of limit points of ordinary probability spaces, in much the same way that irrational numbers can be viewed as limit points (or Dedekind cuts) of benign sets of rational numbers.

[5] Over-protection may also tempt the counterfactual camp; see [Imbens and Rubin, 1995].

Dawid is correct in noting that many problems about the effects of causes can be reinterpreted and solved in non-counterfactual terms. Analogously, some of my colleagues can derive De-Moivre's theorem, $\cos n\theta = Re[(\cos\theta + isin\theta)^n]$, without the use of those mistrustful imaginary numbers. So, should we strike complex analysis from our math books? If we examine the major tangible results in causal inference in the past two decades (e.g., propensity scores, identification conditions, covariate selection, asymptotic bounds) we find that, although these results *could* have been derived without counterfactuals, they simply *were not*. This may not be taken as a coincidence if we ask why it was Eratosthenes that measured the size of the earth and not some Babylonian astronomer, master in black-box prediction. The success of the counterfactual language stems from two ingredients that are necessary for scientific progress in general: (1) the use of modeling languages that are somewhat richer than the ones needed for routine predictions, and, (2) the use of powerful mathematics to filter, rather than muzzle, the untestable queries that such languages tempt us to ask.

Dawid is inviting causality to submit to the Babylonian safeguard of black-box mentality. I dare predict that causality will reject his offer.

References

[Annals, 1934] *Annals of the Royal Statistical Society 1834–1934*, 1934. The Royal Statistical Society, London, page 16.

[Balke and Pearl, 1994] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

[Balke and Pearl, 1997] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, 1997.

[Dawid, 1979] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, 41(1):1–31, 1979.

[Greenland and Robins, 1988] S. Greenland and J. Robins. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128:1185–1197, 1988.

[Imbens and Rubin, 1995] G.W. Imbens and D.R. Rubin. Discussion of 'Causal diagrams for empirical research' by Judea Pearl. *Biometrika*, 82:694–695, 1995.

[Laplace, 1814] P.S. Laplace. *Essai Philosophique sure les Probabilites*. Courcier, New York, 1814. English translation by F.W. Truscott and F.L. Emory, Wiley, NY, 1902.

[Pearl, 1995] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, December 1995.

[Pearl, 2000] J. Pearl. *Causality.* Cambridge University Press, New York, 2000. Forthcoming.

[Pearson, 1911] K. Pearson. *Grammar of Science,* 3rd ed. A. and C. Black Publishers, London, 1911.

[Rosenbaum and Rubin, 1983] P. Rosenbaum and D. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrica*, 70:41–55, 1983.

[Toulmin, 1961] S. Toulmin. *Forecast and Understanding.* University Press, Indiana, 1961.