

1

Statistical causality: Some historical remarks

D.R. Cox

Nuffield College, University of Oxford, UK

1.1 Introduction

Some investigations are essentially descriptive. Others are concerned at least in part with probing the nature of dependences.

Examples of the former are studies to estimate the number of whales in a portion of ocean, to determine the distribution of particle size in a river bed and to find the mortality rates of smokers and of nonsmokers. Examples of the second type of investigation are experiments to compare the effect of different levels of fertilizer on agricultural yield and investigations aimed to understand any apparent differences between the health of smokers and nonsmokers, that is to study whether smoking is the explanation of differences found. Also much, but not all, laboratory work in the natural sciences comes in this category.

Briefly the objectives of the two types are respectively to describe the world and in some sense to understand it. Put slightly more explicitly, in the agricultural field trial the object is essentially to understand how the yield of a plot would differ if this level of fertilizer were used rather than that level or, in smoking studies, how the outcomes of subjects who smoke compare with what the outcomes would have been had they not smoked. These are in some sense studies of causality, even though that word seems to be sparingly used by natural scientists and until recently by statisticians.

Neyman (1923) defined a basis for causal interpretation, using a working and not directly testable assumption of unit-treatment additivity specifying the way a particular experimental unit would respond to various treatments, only one of which could actually be used for a specific unit. In the absence of randomization specific consequences were unclear. His

2 CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS

later more ambitious work (Neyman *et al.*, 1935) is discussed below. The landmark paper on observational studies by Cochran (1965) did deal with causal interpretation and moreover in the discussion Bradford Hill outlined his considerations pointing towards stronger interpretation. Both authors emphasized the difficulties of interpretation in experiments and in observational contexts.

1.2 Key issues

Key issues are first what depth of understanding is involved in claims of causality, that is just what is meant by such a claim. Then there is the crucial matter of the security with which such causality can be established in any specific context. It is clear that defining models aimed at causal interpretation is desirable, but that calling such models causal and finding a good empirical fit to data is in some specific instances far short of establishing causality in a meaningful subject-matter sense. Cox and Wermuth (1996, pp. 219–) suggested *potentially causal* as a general term for such situations. This was criticized as overcautious by Pearl (2000). Indeed, overcaution in research is in general not a good idea. In the present situation, however, especially in a health-related context, caution surely is desirable in the light of the stream of information currently appearing, suggesting supposedly causal and often contradictory interpretations about diet and so forth.

1.3 Rothamsted view

The key principles of experimental design, developed largely at Rothamsted (Fisher, 1926, 1935; Yates, 1938, 1951), contain the essential elements of what is often needed to achieve reasonably secure causal conclusions in an experimental context, even though, as far as I can recall, the word cause itself is rarely if ever used in that period. An attempt to review that work largely in nontechnical terms (Cox, 1958a) centred the discussion on:

- a distinction between on the one hand experimental units and their properties and on the other hand treatments;
- a working assumption of unit-treatment additivity;
- a general restriction of supposedly precision-enhancing adjustments to features measured before randomization;
- a sevenfold classification of types of measurement with the above objectives in mind.

In more modern terminology the third of these points requires conditioning on features prior to the decision on treatment allocation and marginalization over features between treatment and final outcome. It presupposes, for example, the absence of what in a clinical trial context is called *noncompliance* or *nonadherence*. More generally, any intervention in the system between treatment allocation and response should either be independent of the treatment or be reasonably defined as an intrinsic part of the treatment. A possibly apocryphal agricultural example is that of a particular fertilizer combination so successful that the luxuriant growth of crop on the relevant plots attracted birds from afar who obtained their food solely from those plots. Consequently, these plots ended up showing a greatly depressed yield. Thus in

a so-called intention-to-treat analysis the treatment was a failure, a conclusion correct in one sense but misleading both scientifically and in terms of practical implication.

If some measurements had been available plot by plot on the activity of the birds, the situation might have been partly rescued, although by making strong assumptions. In general, causal interpretation of randomized experimental designs may be inhibited by unobserved, uncontrolled and unwanted interventions in the period between implementing the randomization and the measurement of outcome. This consideration is of especial concern if there is an appreciable time gap between randomization and measurement of outcome.

Causal interpretation of observational studies, on the other hand, is handicapped primarily by the possibility of unobserved confounders, that is unobserved explanatory variables. When the effect measured as a relative risk is large some protection against the distortions induced by such unobserved variables is provided by Cornfield's inequality (Cornfield *et al.*, 1959).

1.4 An earlier controversy and its implications

Neyman and colleagues read to the Royal Statistical Society (Neyman *et al.*, 1935) an account of work in which the notion of unit-treatment additivity, stemming from Neyman (1923) and in a sense implicitly underpinning the Rothamsted work on design, was replaced by a much more general assumption in which different experimental units had different treatment effects. A provisional conclusion of the analysis was that the standard estimate of error for a Latin square design is inappropriate. This suggestion led to a vehement denunciation by R.A. Fisher, recorded in appreciable detail in the published version of the discussion. The general issues of the role of randomization were further discussed in the next few years, mostly in *Biometrika*, with contributions from Student, Yates, Neyman and Pearson, and Jeffreys. With the exception of Student's contribution, which emphasized the role of randomization in escaping biases arising from personal judgement, the discussion focused largely on error estimation. Interestingly, the aspect most stressed in current writing, namely the decoupling from the estimation of treatment effects of unobserved systematic explanatory features, is not emphasized.

The specific issue of the Latin square was developed in more detail nearly 20 years later by Kempthorne and his colleagues, confirming (Kempthorne and Wilk, 1957) the bias of the error estimate in the Latin square analysis. Cox (1958b) pointed out the unreality of the null hypothesis being tested, namely that in spite of unexplained variation in treatment effect the null hypothesis considered was that the treatment effects balanced out *exactly* over the finite set of units in a trial. When a more realistic null hypothesis was formulated the biases disappeared, restoring the respectability of the Latin square analysis, as argued by Fisher.

While the status of error estimation in the Latin square may seem a rather specialized or even esoteric matter in the present context, the relevance for current discussions is this. It is common to define an average causal effect in essentially the way that Neyman *et al.* did, that is without any assumption that the effect is identical for all units of study. Of course, if the variation in treatment effect between units of study can be explained in substantive terms that is preferable, but if not, it should in some sense be treated as stochastic, and that affects the assessment of error.

The point is similar in spirit to what is sometimes called the marginalization principle, namely that it rarely makes sense to consider models with nonzero interaction but exactly zero main effects.

4 CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS

1.5 Three versions of causality

In general terms, it is helpful to distinguish three types of statistical causality, all of importance in appropriate contexts (Cox and Wermuth, 2004). The first is that essentially a multiple-regression like analysis shows a dependence not explained away by other appropriate explanatory variables. In a time series context this forms Wiener–Granger causality; see, also, the more general formulation by Schweder (1970). The second definition is in the spirit of the previous section in terms of a real or notional intervention and implies a restriction on the nature of variables to be treated as possibly causal. This is the approach that has received most attention in recent years (Holland, 1986). The third notion requires some evidence-based understanding in terms of the underlying process.

The general principle that causality operates in time and rarely instantaneously has implications for model formulation. Thus if two variables measured as discrete-time series (X_t, Y_t) are such that each component in some sense causes the other a suitable formulation is that (X_{t+1}, Y_{t+1}) depends on (X_t, Y_t) , with a corresponding differential equation form in continuous time. There are appreciable difficulties for subjects like macro-economics, in which data are typically quite heavily aggregated in time (Hoover, 2001).

1.6 Conclusion

It seems clear that the objective of many lines of research is to establish dependencies that are in some sense causal. Models and associated methods of statistical analysis that point towards causality are therefore very appealing and important. However, the circumstances under which causality in a meaningful sense can be inferred from a single study may be relatively restricted, mainly to randomized experiments with clear effects and no possibility of appreciable noncompliance and sometimes to observational studies in which large relative risks are encountered. Really secure establishment of causality is most likely to emerge from qualitative synthesis of the conclusions from different kinds of study. A fine example is the paper on smoking and lung cancer by Cornfield *et al.* (1959), reprinted in 2009. Here large-scale population data, the outcomes of longitudinal prospective and retrospective studies and the results of laboratory work were brought together to provide a powerful assertion of causal effect in the face of some scepticism at the time.

References

- Cochran, W.G. (1965) The planning of human studies of human populations (with discussion). *Journal of the Royal Statistical Society A*, **128**, 234–266.
- Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. and Wynder, E.L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**, 173–203. Reprinted 2009 in *International Journal of Epidemiology*, **38**, 1175–1191.
- Cox, D.R. (1958a) *Planning of Experiments*. New York: John Wiley & Sons, Inc.
- Cox, D.R. (1958b) The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.
- Cox, D.R. and Wermuth, N. (1996) *Multivariate Dependencies*. London: Chapman & Hall.

STATISTICAL CAUSALITY: SOME HISTORICAL REMARKS 5

- Cox, D.R. and Wermuth, N. (2004) Causality: a statistical view. *International Statistical Review*, **72**, 285–305.
- Fisher, R.A. (1926) The arrangement of field experiments. *Journal of the Ministry of Agriculture Great Britain*, **33**, 503–513.
- Fisher, R.A. (1935) *Design of Experiments*. Edinburgh: Oliver and Boyd, and subsequent editions.
- Holland, P.W. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.
- Hoover, K.V. (2001) *Causality in Macroeconomics*. Cambridge: Cambridge University Press.
- Kempthorne, O. and Wilk, M.B. (1957) Nonadditivity in a Latin square design. *Journal of the American Statistical Association*, **52**, 218–236.
- Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. *Roczniki Nauk Rolniczych*, **10**, 1–51. English translation of Section 9 by D.M. Dabrowska and T.P. Speed, *Statistical Science*, **9**, 465–480.
- Neyman, J., Iwaskiewicz, K. and Kolodziejczyk, St. (1935) Statistical problems in agricultural experimentation (with discussion). *Supplement of the Journal of the Royal Statistical Society*, **2**, 107–180.
- Pearl, J. (2000) *Causality*. Cambridge: Cambridge University Press, 2nd edn, 2010.
- Schweder, T. (1970) Composable Markov processes. *Journal Applied Probability*, **7**, 400–410.
- Yates, F. (1938) *Design and Analysis of Factorial Experiments*. Harpenden: Imperial Bureau of Soil Science.
- Yates, F. (1951) Bases logiques de la planification des experiences. *Annals of the Institute of H. Poincaré*, **12**, 97–112.