Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

SOME GENERAL ISSUES AND CHALLENGES FACING STATISTICS

Professor Sir David Cox[1]

It is always a great pleasure to come to Canada. In a way, this is a nostalgic visit for me because the first time I came to this country it was just over 39 years ago and it was to this hotel, where the International Statistical Institute held its session in 1963. I have very happy memories of that visit, both on a personal level and scientifically.

Over the years, I have developed what I thought were successful techniques for giving non-committal titles to papers. I was asked quite a while ago to come and, fairly soon after, for a title. I thought I had given a fairly non-committal title. (I had, in fact, given a totally non-committal abstract, as you'll see if you look in your program.) But I had not quite done it properly because, while I will talk about some general challenges, I really want to talk about one particular challenge which is facing up to the issue of causality in statistics.. In a sense, this issue subsumes some others. Putting it another way, it is facing up to the question: "What do analyses actually mean when we have done them?

If we look at the general statistical literature, the interesting thing about causality is that it is like Charlotte Home's dog. It was interesting because it did not bark in the night. Causality is barely mentioned. This may be because it is thought to be philosophy -and philosophy that was to be kept well away from. Or, it might be thought that causality is a subject matter issue and, of course, that in a certain sense is true. So what this talk is about, in a way, is trying to bring together a bit more the results of statistical analyses and the subject matter interpretation that underlies them.

If we go on from this sort of rather negative attitude but look back to the beginnings of the teaching of statistics, it is quite possible one would come across the slogan "Correlation is not Causation." That might be almost the only time that the word was ever heard.

Correlation is not causality. Causality can be inferred from randomized experiments but not from observational studies. That is the next thing one might hear and that is considerably more than a half-truth- it is perhaps a 95% or 98% truth which we'll come to discuss a bit later on. That is in the general statistical literature.

But the more one goes towards the social sciences or looks in the quantitative literature on sociology or econometrics and certainly epidemiology, then causality is discussed to some extent at least. The words are used in many different ways by different people. Not only in different ways but with different levels of care. I could have mentioned also, in addition to those fields, the artificial intelligence literature, which makes a lot of use of causal models . Some people use the word “causal” pretty freely.

Epidemiologists by contrast, (the best epidemiologists, I think,) are extremely careful in distinguishing between what they call risk factors (which are, for a disease, indicators of association) and claims of causality. So, there are many different definitions and many different levels of use of the word and it is important to understand that in looking at the literature.

I am going to give three definitions of causality which are related to one another but I think it is worth distinguishing between them.

First of all, I am going to talk about what I'll describe as "level one causality". We have something that we think of as a possible cause and, just for simplicity of exposition, I am going to suppose that it is binary. We have individuals worth studying and, on those individuals, we have a response. We might have children being educated by one of two different methods and some measure of success in their subject; We might have criminals about to be either dealt with by one or two alternative methods of punishment and some measure, R, of recidivism; or clinical trial patients

[1] Nuffield College, Oxford, United Kingdom

being treated by one of two possible methods, C0 and C1, and the response is some measure of survival or whatever. The notion here is that any individual might conceptually be either C0 or C1. Even if a particular child is taught by method C0, that child could conceptually have been taught by the other method C1.

We talk about causality, in this context, if the response that an individual would show under one method, C1, differs systematically from the response that that individual would have shown under C0, other things being equal and I'll italicize that expression "other things being equal". In other words, this is immediately relevant, for example, to the randomized clinical trial where a patient might get one of two possible treatments. The patient actually gets one of them but could clearly have got the other if the randomization had turned out differently. In the case of the criminological study, a criminal may be dealt with by some system of restorative justice, for example, or not. Maybe the criminal is dealt with by restorative justice but it could have been otherwise. So the world could have been different from how it actually is.

Causality is the notion that other things being equal (and that is a key proviso), there is a systematic difference between those two responses. Being a bit more specific, we can talk about two potential responses, R0 and R1, on the particular individuals, depending on which of C0 or C1 applies. We are allowed to observe one of those, but only one of them. The other is what philosophers would call a "counterfactual". It is something that we can conceive of but cannot actually directly observe. Then, we have to formalize the notion that R1 is systematically different from R0. I'll talk about R1 being greater than R0 just to be specific. So, one might say that R1 is greater than R0 for all individuals. Every possible individual does better under C1 than that individual does under C0, let us say. An even stronger version would be to say that the difference between the R's is a constant. A more realistic version, in many contexts, would be to say that R1 is stochastically greater than R0. That is to say, the cumulative distribution function of R1 is shifted to the right compared with the cumulative distribution function of R0. That is much less than saying that for each individual, the inequality holds. An even weaker version would be to say that the expected values differ. On these last two, because they involve some notion of probability, they implicitly, or maybe explicitly, involve the notion of a target population to which the statement applies.

If we go on from the discussion of this definition, an important point, I think, is that it places a restriction on the admissible variables that could be thought of as causal because it amounts to saying: "OK, we have observed somebody at C1 but that individual might have been a C0." The world might have been different from how it actually is.

I was taught an example of this, almost at the very beginning of my career in research, when I was working with a physicist who told me that you could not take the passage of time as a cause. David Binder might remember that when I was younger I had dark brown hair. I barely remember, but David probably does. It is not legitimate under these terms to say passage of time has turned my hair gray because what would it mean under this definition? It would mean, starting with my brown hair long time ago, passage of time hasn't happened but we have got to today and my hair has gone gray. That is a sort of meaningless statement because it is not conceptually possible to talk about getting from the past to the present without time passing.

We can say the stress of preparing lectures for conferences has turned my hair gray. The issue is not whether this statement is true but whether it is meaningful. It is meaningful because time could have passed. I could never have prepared a lecture for a conference. I could have spent the time in idle pleasure, like say reading the Annals of Statistics or something of that sort, and maybe my hair would not have gone gray. I do not think this statement is actually true but it is meaningful, that is the point. This does have a bearing on issues of interpretation.

Gender differences, for example, will rarely be potentially causal in this sense. The same sort of reason to say: "Here's this woman. What would the response have been, other things being equal, if that woman had been a man"? In most contexts it does not make sense. In an employment discrimination context it might very well make sense, but not in general.

So, it places a restriction on the admissible causes. The approach has an implicit, but not explicit, technological slant to it. It has the notion that we are interested in policy decisions about how to treat patients or, with economic policies, whether raising bank rates or interest rates has an effect on the economy and what it might be, issues of that sort. It has a kind of almost decision-making flavour to it, but that is not necessary. We could say a geneticist could very well say the presence of a particular combination of alleles on the genome causes a particular disease.

(They do say things like this.) Meaning that if, for a particular individual, the haplotype had been different from how it actually is, then the outcome would have been different. That does not necessarily have a decision-making slant to it but there is a sort of a decision-making focus to it.

There will be available a list of references attached to this subject, more than I am actually referring to in this talk. The best place to start reading about this, if it is a new subject to you is a review paper, a discussion paper, in the Journal of the American Statistical Association, quite a few years ago now, by Paul Holland. The general notion that is involved in this really goes back to the beginnings of the theory of the design of experiments in the work of Neyman and Fisher. Cochran wrote a very important paper on the analysis and interpretation of observational studies which brings out many of the ideas that are involved, but it is Don Rubin, particularly, who worked with Cochran, of course, who has developed this idea and applied it in all sorts of social science contexts. Jamie Robins has worked more in an epidemiological setting but has used rather similar ideas.

The notion of a counter-factual leaves many people somewhat uneasy. If you take the view that the only defensible things that really matter are the things that you can actually directly observe, then you can discard anything like variables and anything of that sort. If you have to be able to test directly what you observe, then you do not like the notion of a counter-factual because of its very nature. Phil David has written a very interesting paper criticizing this view of causality from that perspective and again there is an interesting discussion in the Journal of American Statistical Association which, by and large, is defensive of this view of causality. And so, it is about the notion that if the world were different from how it actually is, than in certain respects C, then the response would be different from how it actually is. The notion is that if we intervene on the C we will change the world, hopefully in a positive sense. I call that level one causality.

Level two causality is, in a way, something quite different, superficially at least. It is connected with the notion of understanding that we observe some regularity, some statistical regularity, maybe even a causal relation, in the sense that I have just described. And you could say that the reason that this is happening is because of some processes going on that you at least partially understand. So it is an evidence-based understanding of the generating process.

Perhaps, an example would be something like this. You might collect a lot of data on thunderstorms, their duration, their size, the number of flashes of lightning, the position over which they are centered, so on and so forth, and you might perhaps deduce some statistical regularities out of that that were very clear and very stable and interesting. Almost inevitably one would say: "But what is the physics underlying this? How can all this be related to the basic equations of physics, of fluid dynamics and of electromagnetism? And then one would understand the cause of these relations, in that sense, caused in the sense of understanding.

My colleague in Oxford, a sociologist, John Goldthorpe, has argued that this is the right definition, or the most fruitful definition, also for sociology, going in a way extreme from physics. The statistical regularities that the sociologists establish should, if possible, be explained somehow either in terms of individual psychology or in terms of something like rational action theory. So it is a notion of understanding. It is inevitably provisional and, by its very nature, it involves studies at various levels.

I have already mentioned Cochran's very important paper on observational studies in 1965 and in it he quotes a conversation with R.A. Fisher in which he asked Fisher "What could be done to make observational studies and observational analyses more causal?" Fisher made this characteristically enigmatic reply "Make your theories elaborate", of course, quite counter to any notion of Ockham's razor and simplicity. Make your theories elaborate. What Fisher meant by that, Cochran explains, was: Bring together all sorts of different information of different kinds, different kinds of studies, different levels of understanding into a coherent scheme and, in that way, causality is more closely established.

Bertrand Russell, philosopher, (thinking as philosophers of science have very often done, at least the earlier ones, concentrating very largely on equating science to physics,) initially dismissed at least the notion of causality in this sense as unnecessary. I think, somewhere, he wrote, it is like the monarchy. It had outlived its usefulness and it only lingered on because of the mistaken belief that it was harmless. He changed his notions about causality, I think, but whether he changed them about the monarchy (or not) I do not actually know. Anyway, make your theories elaborate.

So, superficially at least, that is quite a different notion of causality, basically about understanding. My impression is that that is the notion that most natural scientists would think of if they used the word. By and large, they do not use the word very much, but I think, if pressed, that is what they mean. I do stress this expression, "evidence-based", because, after all, for any phenomenon, almost in any field, the clever human mind can think of an explanation. However bizarre the phenomenon might be, someone will cook up a possible explanation of it. That may be very useful as a first step but the explanation, in this sense at least, has to be evidence-based. Provisional, yes, but evidence-based. That is two definitions of causality.

I want to give a third definition. I am not being dismissive in this, but I have called this level zero causality. This is what sometimes is called probabilistic causality. You may have noticed that although I have talked for a quarter of an hour I have hardly used the term probability. I used the term probability. I did talk about distribution functions at one point. This is probabilistic causality. The notion is that the distribution function of our response under C1 is less than or equal to that under C0. The distribution function shifted from the true distribution with strict inequality on some set of positive probability. So the distribution functions are separated, There is a stochastic ordering between the two distributions. So that just says that C and R are associated. Because that is nothing like enough for a reasonable definition of causality, one has to go on from that. There has to be some external information to say that it is C that precedes R and not visa versa. So the direction of the causality needs to be established somehow, for example, by temporal ordering or by hypothesis.

My colleague Nanny Wermuth has a very interesting study in medical sociology, basically, of the relation between diabetics' knowledge of their disease and their success at controlling it and showing a positive relationship for people who know more about the disease to doing better at controlling it. The tentative explanation of that was that the knowledge causes, in some sense, a better control. That is a working hypothesis. The data were cross-sectional and so it is certainly conceivable that it could go the other way round - that the people who are doing well at controlling their disease are encouraged to learn more about it. It could go in that direction. It was a working hypothesis that the causality went that way.

Anyway, so we need an ordering and then very importantly, we need to assure ourselves that there are no allowable alternative explanations. That, further conditioning, by an allowable variable, whatever that might mean, will not destroy this relationship.

The issue of what is an allowable variable in this context, being a bit more statistical for a moment, is what variables should you put in to your regression equation in which you are trying to study the relation between R and C?

The issue of what are allowable variables is closely connected with the notion, in the level one causality, of other things being equal. That definition, I think, goes back to I.J. Good and is set out most systematically by the American philosopher of science Patrick Suppes. In econometrics, this notion is applied to time series. There is a fairly obvious generalization of this notion when R and C are both time series. That is, in econometrics, called "Granger causality". We know the mathematician had a very similar notion in a general context.

So these are three notions of causality. Three, if you like, definitions of causality. The world would be different if C changed from C0 to C1 as regard to the response R, other things being equal. We understand, in some sense, why that is, which is level two. Level zero is a more directly probabilistic notion. We can establish a statistical relationship between C and R. We are satisfied the dependence goes from C to R and not visa versa and we are satisfied that we cannot explain it away somehow. Those are definitions of causality - particularly level zero and level one. They are definitions. The issue of how we actually might verify this definition in a particular case- how we might satisfy ourselves that causality really obtains in a particular situation , in a way, is the essence of the problem. “That is much more difficult” one might say. The last definition, certainly, outside the realm of randomized experiments, is not directly implementable because we have to satisfy ourselves that there is no possible alternative explanation. It is always possible in an observational study that there is an alternative explanation and we just have not measured the appropriate variables to assess it. So one might say that many of these things are sort of targets at which to aim but they are not directly achievable.

I do not know how much time to spend on this because I do not know to what extent this helps. This is just an example that came up in the newspapers a few weeks ago- a study - I must confess I have not looked at the details at all, so I am talking rather hypothetically-a study that said taking exercise reduces the risk of certain kinds of cancer.

What might that mean if it is to be interpreted causally? In the first definition, it would say: Here are people who have taken a lot of exercise. Here, let us say, is a group of people who have taken a lot of exercise. Had they not taken exercise but otherwise remained the same, in some sense, then more of them would have gotten these particular cancers and visa versa for those who did not take exercise.

That is a fairly definite definition. The issue is: how would one actually establish that this was true? If this was being done on animals, on mice should we say, we can conceive of a randomized experiment in which an appropriate number of mice were randomly divided into two groups. One, as it were, forced to take a lot of movement and the other group constrained in their movement and perhaps, too, one would use transgenic mice and perhaps irradiate them to increase the rate of cancers. But one might hope to show some difference between the two groups. There can be no other explanation. If there is a big difference, either the play of chance has fooled you, you have been very, very, unlucky, or it is to do with the amount of exercise because there is not any other explanation. The two groups are otherwise comparable.

But, with an observational study, and trying to apply the level zero definition, you would have to be able to say that the people who take exercise and the people who do not - they have the same age distribution, the same gender distribution, and the same socio-economic class, and so forth. When was the exercise taken? Could it be that the people in which cancer has been initiated, but not promoted, are inhibited from taking exercise and that the reason that they do not take exercise is because of the initiation of cancer. There is a minefield, obviously, of difficulties of interpretation in implementing that definition.

The level two definition of understanding would involve trying to show that there were biochemical, physiological processes going on, encouraged by taking exercise, that, in fact, inhibited either the initiation or promotion of cancer. That would be trying to understand the processes that underlie what you observed. So establishing such a thing empirically, either in the level zero or level one definition, is clearly very difficult in the non-experimental situation.

Of course, there are further complications. George Bernard Shaw, I think, wrote in the preface of one of his plays something like "Exercise. If you're fit, you do not need it and, if you're unfit, it is dangerous". In other words, he was suggesting an interaction effect between an initial health status and the effect of taking exercise. That would be a further level of complication in this.

I have already emphasized, I think, that almost all that I have said so far is about what causality might mean and not very much about, in many ways, a much more pressing issue. How, both in a survey context and more generally, do we set about establishing anything at all like causality? Very important work has been done recently coming partly from the philosophical school at Carnegie Mellon, from Glymour and others, and very particularly also from an influential book by Judith Pearl from UCLA. I want to, in a way, pass to that and the relation between what Judith Pearl has proposed and what I would think of as standard statistical thinking.

To do that I am going to go to a very idealized situation where we have a study in which we have basically got four kinds of variable.

We have got a variable that we think might be a cause, C, and we have got a variable that we think of as appropriate response, R. There are what I call a background variable, B, and an intermediate variable, I. B and I, in particular, could be highly multi-dimensional and, I suppose, so could R, for that matter. C, I am thinking of as one-dimensional for the moment. The first assumption that is made in all this work, is that this system forms what they call a directed acyclic graph. In more concrete terms, what that means is that for any pair of these variables we know the direction. They may or may not be associated, but, if they are associated, we know the direction. So this would be most tangible if B, C, I and R occurred in a longitudinal sense in time. (Just appealing to the notion that the future is a response to the present, so to speak, and not visa versa.)

The idea of the directed acyclic graph is that we have directed edges between these nodes and missing edges represent some kind of conditional independence. I am not going to go into the details of that, so much, but just to emphasize that the direction of dependence is assumed known.

We are interested in whether C affects R. By the first definition of C, what we want to be able to say is, if we were to change C from C0 to C1, or if we were to change C by unit amounts, if it was an exposure measure on a continuous scale, what will happen to R? Superficially, what we have to do is to run a regression of R on something or other and examine the coefficient of C because regression coefficients, in the least squares sense at least, tell us the change in R per unit change in C. What regression should we run?

What Pearl, in particular, does is the following: First of all, we ignore I, I is marginalized over it and anything that occurs that is intermediate - in the sense that I is a response to C but it is also explanatory to R. Perhaps I should give a simple example. Patient's blood pressure. At the beginning of the study, C is some medication at different levels. I is the blood pressure after three months and R is death from heart disease within five years, yes or no. So we'd have a logistic regression of the response on the previous variables and the blood pressure after three months. I, having been influenced probably by the treatment, would not be included in the study of the direct effect of C. So examine the conditional distribution of R given C and B.

In a linear system, that is to say a system where either everything is multivariate normal or reasonably characterized by least squares regression equations, there is a relationship between the total regression coefficient of R on C, the partial regression coefficient of R on C given B and what that says is that the total relation of C to R is built up of the direct path from C to R, plus an indirect path as shown in this diagram. This idea goes back to the geneticists who wrote in the 1920s - the late 1910s actually. Another person who produced a brilliant idea in his doctoral, or pre-doctoral time - as we heard yesterday - will have that relationship.

What Pearl does is to marginalize over C. He considers the relation between R and C given B, and then he integrates over the distribution of B but it is not the conditional distribution of B given C, it is the marginal distribution of B because, having this notional change of C, we are interested in what is the effect of manipulating C. When we manipulate C, we do not affect B because doing something today does not affect the world yesterday. Pearl makes great play of what he calls the causal calculus, as distinct from the probability calculus, but what it really amounts to is that he is integrating over the distribution of B that is marginal rather than conditional.

But statisticians would not do that anyway, I think. They would regress. They would, I hope, ignore I for the time being; regress R on C and B; look for interactions between the effect of B on C (The effect of C may be different from men from what it is for women); and look for interactions. If none seem to be present, hey would attempt to find a specification of the effect of C that was invariant with respect to the background variable.

What is the relation of this with the other definitions of causality? In an observational study and, for that matter, in an experiment as well, there may very well be unobserved background variables - variables that are partly explanatory of C, but which also are unobserved - either because we have not bothered to observe them for reasons of economy or because we may not even quite know what they are. In the case of exercise, it could very well be some sort of variable, maybe even a latent variable, general health status at the beginning of the period in question, for instance.

We can apply the same argument again. We want to regress R on C, B and U. We cannot do that because we do not know what U is. So what we have to do is to regress R on C and B and ask ourselves the question: Does it really matter that we have not observed the background variable? In the square sense, and it turns out more generally, the condition that is required is that this term here is either 0 or at least very small. There are two ways that that can happen. Either this is 0 which says that R is independent of U, given C and B, so in this picture here where U, the unobserved variable, affects C, it affects B and it affects R. In fact, there is no edge in the graph from C to R. So that is just saying, sort of obviously in a way, that if the unobserved variable does not have any effect on the response other than is accounted for, what we have got in the model already, it does not matter, at least in the linear sense. (There is a mistake in the original transparency, it should say C is independent of U given B or this edge is missing.) So, it may be that the background variable affects the response but it has no effect on the cause except via things that we have already observed.

In a randomized experiment, that is satisfied by definition because we choose C, in a way. It may depend upon observed features, on strata and all the rest of it. I am talking about random allocation and not random selection in the sampling sense. In a randomized experiment, this edge is missing by definition. So, in this formalism, the role of the randomized experiment is established by the argument that edges like that, are, by definition missing. In an observational study, we have to try, if we want to interpret the regression coefficient of R on C given B causally - that is to say what the influence of changing C would actually be. If we want to interpret it causally, we have to satisfy ourselves that either this edge is missing or the one or other of these edges is in some sense small. Of course, that is difficult.

Just a few general consequences. I expressed all that in terms of these graph theory representations. You do not have to do it that way. You can do it all in terms of formulae but maybe the graphs may help. A much more mathematical point, just very quickly, is that it is known that, in principle, all results in graph theory -I am talking about graph theory in the algebraic sense of the term - of collections of nodes with directed and undirected edges between them, and some edges missing- can always be related to results in matrix algebra basically, specifying the presence of edges by incidence matrices. Professor Wermuth and I have recently shown, in some detail, the relation between these graph-theory-like maneuvers based on so-called separation theorems and the underlying matrix algebra.

Randomization, I have talked about it. How can we deal with this assumption that we need? All I am saying - in a way you may feel I am making a great undue, should I say, very heavy weather of this - is that the effect that interpreting regression coefficients causally(I am using regression coefficients in a general sense) requires observational data and requires assumption about the effect of unobserved variables. One way of assessing this is by sensitivity analysis. Paul Rosenbaum's book on observational studies (the second edition just appeared) goes into this very nicely. I will not say anything about the last point as I am slightly running out of time.

I have described, in outline, how Pearl and others (and a large part of the literature on this) deal with these problems by directed acyclic graphs. I think, to deal with issues more realistically, what we have to do is to be, unfortunately, quite a bit more elaborate about the nature of the variables that we measure on each study. (I am going to go through this fairly quickly.) This is, in a way, just a formalization of relatively obvious ideas. For each study individual, we have got intrinsic variables that define who that individual is, so to speak. We have got background explanatory variables that we do not think of as possibly causal but which are important to record. We have variables that might be possibly thought of as causal. We have got intermediate variables and then we have got response variables. All these things, in general, will be multi-dimensional, of course, in the situation of realistic complexity. What we then have to do is to think about the relation between any pair of variables. I think what we have to do is to say that for any pair of variables, either one of them is potentially explanatory to the other or they are to be treated on a equal footing. So you may have measured diastolic and systolic blood pressure on individuals. I am not saying it is a subject matter issue. -There might be contexts where that did make sense, but in general it would not. - So we describe the variables as either on an equal footing for any pair of them or say one is explanatory to the other preferably based on temporal ordering but, if not, based on a working hypothesis. Then, what we can do is to get a generalization of the directed acyclic graph in which we have the variables arranged in boxes -like that- in such a way that the edges between any two boxes - the variables in two different boxes- go from left to right but the edges within a box are undirected. So one needs a development of the theory to deal with so-called chain block graphs. That is, I think, almost essential for realistic progress on these things.

Going back to a more practical issue. In an observational study - a non-randomized study- when we find a potential cause C that seems to be positively related to a response R, (when we can reasonably assume that that relation is causal in the sense, let us say, of the level one definition). If we changed C, other things being equal, R would change.

Bradford Hill, a medical statistician, suggested a number of guidelines where this might be true. These guidelines - although he suggested them in the context of environmental epidemiology - are quite broadly applicable. He emphasized that they are guidelines, not criteria, and they are certainly not necessary and sufficient conditions. They can be related, to some extent, to the semi-quantitative discussion I gave a moment ago. If the effect that we find is large, -the issue is we find it is large.- we have looked at a possible dependence on alternative background variables. We have not explained away the effect that way. Is it possible that there is an unobserved variable that is the real explanation of what we see? If the effect is large it is less likely that there is an unobserved variable that we do not know about that is the real explanation. In the mathematics, it is just that if the primary term that we are studying is large, it is less likely that this term is its real explanation.

If the effect is monotonic in a measure of exposure - if we have got instead of a binary potential cause, a quantitative level of exposure say to some environmental hazard, - it is more likely to be causal. If the effect is reproduced in different studies and for different types of individuals, - in other words, if there is an absence of a qualitative interaction with background variables - then it is more likely to be causal.

If there is a subject matter explanation, - modifying Bradford Hill's version slightly, I would think it would be better to say, when the subject matter explanation is preferably obtained a priori because I remarked earlier almost everything can be given a subject matter explanation with sufficient ingenuity - and if the explanation either follows from absolutely cast-iron principles, - like say the principles of quantum mechanics or classical mathematical physics- OK. But if it is just produced speculatively with hindsight, it may be true and it may useful to do that, but it is less convincing.

In the context of what Bradford Hill called a natural experiment, -that often is some sort of catastrophe or major event that is so extreme that it is unlikely that its consequences are anything other than the direct cause of whatever happened. A standard example of that is the study of the health of the atom bomb survivors in Hiroshima. - the relation between their health exposure and the dose of radiation to which they were exposed is strong and it merely is pretty inconceivable that it is other than a causal reaction to that event.

The last one is much more controversial and some epidemiologists strongly challenge this. Bradford Hill said if the relationship is specific - in the disease context what he meant was that if the exposure - whatever it might be - induced cancer at one site but not at others - he was thinking, of course, of smoking and lung cancer - things like that. It affected one site and was very specific in its action. - it was more likely to be causal. Whereas, with lots of different response measures, it was less likely to be causal. That is a controversial issue and many epidemiologists would not go along with that.

I have time to just say, I think, a couple of other things. I introduced those intermediate variables in the picture that we had - intermediate between the potential cause and the response - and then immediately threw them away. The point of putting them in was to partly establish the point. It is not an issue of regressing the response on everything that you have measured. Some things that are after, in some sense the potential cause, should not be included. So why do we think about them at all?

There are a number of reasons. They may be monitoring what actually happens in the process. They may be indicators of something that is independent of the process that we are looking at that disturbs what we are studying. They may be useful as surrogate responses. There is a whole issue as to what constitutes a legitimate surrogate response. (The situations where it is difficult or expensive or time-consuming to measure the response that you're really interested in.) They may be helpful in establishing causal paths between the cause C and the response R i.e. moving us somewhere towards level two causality.

To illustrate the second thing I want to tell you, let me tell you a not totally apocryphal story about a randomized experiment. In a classical agricultural field trial with treatments randomly allocated to plots, one fertilizer produced a very beautiful crop that grew luxuriantly and birds - Remember it is a randomized trial with every possible precaution for the avoidance of birds. - birds from many kilometres around descend on these plots and eat all the product. So when the final response is measured those plots have a very low yield. Did this magic fertilizer cause a depression in the yield? In one sense it did, you see because those plots that received that magic treatment did worse than they would have done if they had had some other treatment. The level one causality is satisfied.

Level two causality has all sorts of randomized trials you cannot explain. There is no background variable that explains this. Yet that would be a very misleading conclusion to draw, either as a scientist, because it would lead you to a total misunderstanding of what was going on, or as a technologist. If you were a farmer you would say: “What fertilizer am I going to put on my crop?” (I am talking about the days when growing more food was thought to be a good thing.) What fertilizer should I put on my plot? Not the magic one? That would be the wrong conclusion because if you use the magic fertilizer on large areas, then the birds would have eaten their fill over the area and the nice high quality, high quantity yield would have been plain.

The conventional dictum is anything that happens after the instant of randomization would be potential cause. But the point is that is not universally true. In a way it illustrates one of the things that David said he learnt from the terrible examples in my book with David Hinkley of things that can go wrong. It is an example of a principle. Do not ignore the variables between the cause and response. But it is an example. If you take this principle to the bitter end then you come up with silly answers. So, intermediate variables serve that purpose.

Somewhere I have got a summary of all this. What is the summary of what I have just said? In a sense, that the search for causation is ubiquitous. If we want to understand what our results mean, in some sense or other, we are looking for causation. The word can be used in various ways. I want to emphasize being very cautious about the use of the word. You will in some areas of applied statistics find models called causal models simply because one would like them to be causal. I am implicitly suggesting a more cautious approach. Look for analyses that at least point in the direction of causality preferably in all the three senses that I have mentioned. There is, in a sense, somewhat of - not a conflict but - a tension between that and what I would regard as the empirical ethos of statistics. The empirical ethos of statistics is a pompous way of saying: Stick close to the data. What do the data tell us?

There is a tension there that we should stick to what the data tells us and try and let the data point us in the direction of interpretations that have subject matter weight to them.

So the summary of the whole lecture, in a way, is to be either cautiously optimistic or optimistically cautious, I am not quite sure which.

Thank you very much.

SLIDES FROM PROFESSOR COX'S PRESENTATION

Causality (Slide # 1)

Statistical attitudes to causality

- Its best left to the philosophers, i.e. ignored
- Correlation is not causation
- Causality can be inferred from randomized experiments but not from observational studies
- But appreciable interest in social statistics, econometric and epidemiological circles and there is more general statistical involvement recently

Level one definition (slide # 2)

Potential cause C, taken binary for simplicity, C_0, C_1.

Response R.

For each individual only one C and associated R can be observed.

Response of an individual under C_1 systematically different from the response that would have been observed on that individual under C_0 *other things being equal*.

- potential responses, R_0, R_1
- one is counterfactual
- various ways of formalizing systematic difference, for example
  - $R_1 > R_0$ for all individuals
  - $R_1 – R_0 = \Delta$
  - R_1 stochastically greater than R_0
  - $E(R_1) > E(R_0)$
  - Last two involve an explicit or implicit study population
- places a restriction on admissible C
- approach has a (not inevitable!) technological slant
- Holland (1986); Neyman and Fisher; Cochran (1965); Rubin (1974); Robins (1997), Dawid (2000).

Level two definition (slide # 3)

Evidence-based understanding of generating process

- involves studies at various levels
- inevitably provisional
- Fisher's dictum: Make your theories elaborate
- Bertrand Russell, Goldthrope (2000)

Level zero definition (slide # 4)

Directly probabilistic definition

$$P\ (R \leq r \mid C_1) \leq P\ (R \leq r \mid C_0) \qquad (1)$$

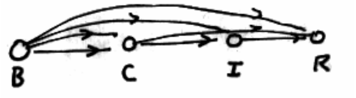
Plus condition that

- C precedes R temporally or otherwise
- There is no further conditioning by an *allowable variable* that will destroy (1)
- I.J. Good, Suppes, Wiener, Granger

An example (slide # 5)

Taking of exercise.

Simple special case (slide # 6)



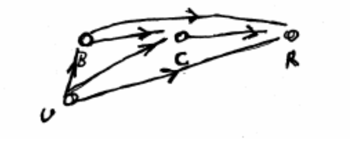
For causal assessment, marginalize over *I* and condition on *B* and examine how *R* depends on *C* in $f_{R|C,\, B}$.

Linear system

$$\beta_{RC} = \beta_{RC.B} + \beta_{RB.C}\beta_{BC}.$$

If there is an unobserved component denoted by *U*, marginalizing over it occurs and has no effect on the slope of the relation between *R* and *C* if

$R \perp\!\!\!\perp U \mid C, B$ or $C \perp\!\!\!\perp U, B$.



Some consequences (slide # 7)

Graphical interpretation
Matrices and graphs
Randomization
Assumption plus sensitivity analysis
Preservation of monotonicity of relation

A general formulation (slide # 8)

For each study individual classify the variables recorded as

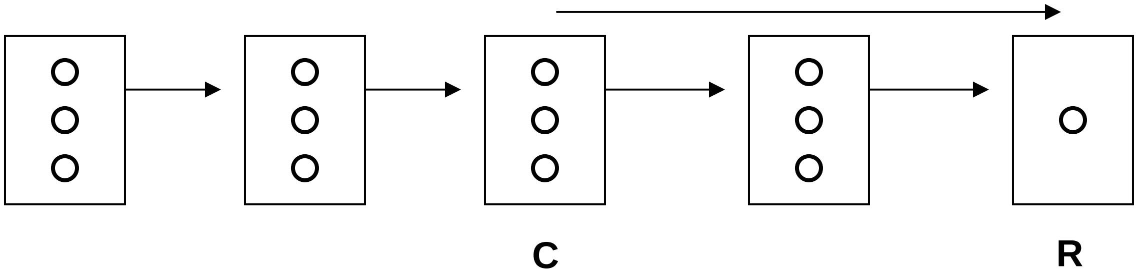
- Intrinsic variables that define the individual under study
- Background explanatory variables that are not to be treated as potentially causal
- Potential causal variables
- Intermediate variables
- Response variables

Distinctions highly context dependent

Assume any pair of variables such that

- one is explanatory to the other
  - Based on temporal ordering
  - Working hypothesis
- variables on equal footing

Graphical representation (slide # 9)



Directed edges between blocks
Undirected edges within blocks
Absence of edge signifies conditional independence

Bradford Hill’s guidelines (slide # 10)

Bradford Hill (1965) suggested guidelines (not criteria) in the context of environmental epidemiology suggesting when an association found in an observational study is likely to be causal.

- when the effect is large
- when the effect is monotonic in “dose”
- when the effect is reproduced in different studies and for different types of individual
- when there is a subject-matter explanation
- in the context of a “natural experiment”
- when the effect of *C* is specific

Summary (slide # 11)

- search for causation ubiquitous
- various uses of word *causal*
- analyses that are potentially causal
- relation to empirical aethos of statistics

Moral (slide # 12)

Optimistic caution

Or maybe better

Cautious optimism

References (slide # 13)

Box, G.E.P. (1966). Use and abuse of regression. *Technometrics* 8, 625-629.

Bradford Hill, A. (1965). The environment and disease: association or causation. *Proc. R. Soc. Medicine* 58, 295-300.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press.

Cochran, W.G. (1938). The omission or addition of an independent variable in multiple linear regression*.* *Suppl. J.R. Statist. Soc*. 5, 171-176.

Cochran, W.G. (1965). The planning of observational studies in human population (with discussion). *J.R. Statist. Soc.* A 128, 234-265.

Cox, D.R. (1958). *Planning of experiments.* New York: Wiley.

Cox, D.R. (1968). Notes on some aspects of regression. *J.R. Statist. Soc*. A 131, 147-174.

Cox, D.R. (1984). Present position and future developments, some personal views: Design of experiments and regression. *J.R. Statist. Soc*. A 147, 306-315.

Cox, D.R. (1992). Causality: some statistical aspects. *J.R. Statist. Soc*. A 155, 291-301.

Cox, D.R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life. *Lifetime data analysis* 5, 307-314.

Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science* 8, 204-283.

Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies*. London: Chapman and Hall.

Cox, D.R. and Wermuth, N. (2002). On the qualitative effect on dependencies of marginalization. Submitted.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *J. Amer. Statist. Assoc*. 95, 407-448.

Dawid, A.P. (2002). Influence diagrams for causal modeling and inference*. Int. Statist. Rev*. 70, 161-189.

Doll, R. (2003). Can epidemiological methods establish causality? Fisher memorial lecture. October 2001, in preparation.

Fisher, R.A. (1926). The arrangement of field experiments. *J. Ministry of Agric*. 33, 503-513.

Fisher, R.A. (1935). *Design of experiments*. Edinburgh: Oliver and Boyd. And subsequent editions.

Goldthorpe, J. (2000). *Causation, statistics and sociology*. Chapter 7 in On sociology. Oxford University Press.

Good, I.J. (1961). A causal calculus: I. Brit. *J. Philosoph. Sci.* 11, 305-318.

Granger, C.W.J. (1969). Investigating causal models by econometric models and cross-spectral methods. *Econometrica* 37, 424-438.

Holland, P.W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* 81, 945-970.

Hoover, K.D. (2002). *Causality in microeconomics*. Cambridge University Press.
Lauritzen, S.L. (1996). *Graphical models*. Oxford University Press.

Lauritzen, S.L. (2000). Causal inference from graphical models. In *Complex stochastic systems*. Pp. 63-107. eds O.E. Barndorff-Nielsen et al. London: Chapman and Hall.

Lindley, D.V. (2002). Seeing and doing: the concept of causation (with discussion). *Int. Statist. Rev*. 70, 191-214.

McKim, V.R. and Turner, S.P., eds (1997). *Causality in crisis?* University of Notre Dame Press.

Pearl, J. (2000). Causality. Cambridge University Press.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 8, 431-440.

Robins, J. (1997). Causal inference in complex longitudinal data. In *Latent variable modeling with applications to causality*. Pp. 69-117, ed. M. Berkane. New York: Springer.

Rosenbaum, P.R. (2002). *Observational studies.* Second ed. New York: Springer.

Rothman, K. and Greenland, S., eds (1999). *Modern epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven.

Rubin, D.B. (1974). Estimating causal effect of treatments in randomized and nonrandomized studies. *J. Educational Psychol*. 66, 688-701.

Scheines, R. (1997). An introduction to causal inference. Pp. 185-199. In *Causality in crisis?* Editors V.R. McKim and S.P. Turner. University of Notre Dame Press.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction and search*. New York: Springer.

Stone, R. (1993). The assumptions on which causal inference rest. *J.R. Statist. Soc*. B 55, 455-466.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: N. Holland.