Arnold Faden
Sept. 9, 1992

CAUSALITY AND ECONOMETRICS IN FOCUS

I. The Knowledge Industry, Methodology and Statistics

(This section is reprinted from my handout of April 21, 1992, entitled The Future of Economic Theory: The FOCUS Paradigm).

Knowledge production is one industry among others. Economics is one of its branches. The industry is organized in terms of professions. The output consists of journals, books, papers, and, ultimately, of changes in people's cognitive states.

Most of the markets for this industry are missing, so that, as in all cases of this sort, a system of imputed valuations has arisen, involving some incomplete degree of consensus among the workers in each branch: the "seminality" of ideas, or, more crudely, the number of pages published. (A rough model of the knowledge industry would have its workers maximize publications, just as politicians are supposed to maximize votes, bureaucrats the size of their budgets, etc.)

Statistics is that branch of the knowledge industry which provides canons of acceptable treatment of empirical data for the other branches. (The other branches invariably modify these canons for their own purposes; econometrics is the modification of statistics adapted to economics).

As in any industry with imperfect markets, there are distorted incentives and inefficiencies in the knowledge industry in general, and in its economic branch in particular. We will concentrate on the econometric canon.

Non-econometricians, including journal editors, do not have the time or inclination to examine the foundations of inference critically. What they need is a "cookbook" of fairly simple procedures and conceptual tools for handling data -- e.g. ordinary significance tests at conventional significance levels.

If one does go more deeply into the foundations of statistics, one finds that almost everything is unsettled, up to and including the concept of probability itself. Standard statistical procedures have been subjected to withering criticism -- e.g. by Lindley, Savage, Berger, Jaynes, Leamer, et. al. All of these criticisms have been in terms of the internal coherence (or incoherence) of procedures, however. We propose to go beyond this and think of statistics in terms of the role it plays in the knowledge industry.

In effect, this approach endogenizes statistics (and its sub-branch, econometrics) into economic theory. The value of statistical procedures reflects the expected real social value of the "theses" they help produce.

This requires taking account not only of the internal coherence of procedures, but of the social and psychic context in which they operate -- the bounded rationality (limited informational capacity) of the human mind, and the behavior of writers, readers and editors in the knowledge industry.

Internal coherence alone leads to Bayesian inference. To grasp the issues it is necessary but not sufficient to understand Bayesian inference: other considerations, such as simplicity of models and the need for "objective" consensus, also enter.

II. The Value of a Model

At any time a person has a "model" of the world in his mind (Fig. 1). The term "model" is not standard. Synonyms might be "image" (Boulding), "lifeworld" (Husserl), "behavioral world" (Koffka), "umwelt" (Uexkull), or even just "world" - e.g. in "Christina's World". In general, a model is subjective, partial, involves uncertainty as well as error and distortion. A more formal treatment is attempted later.

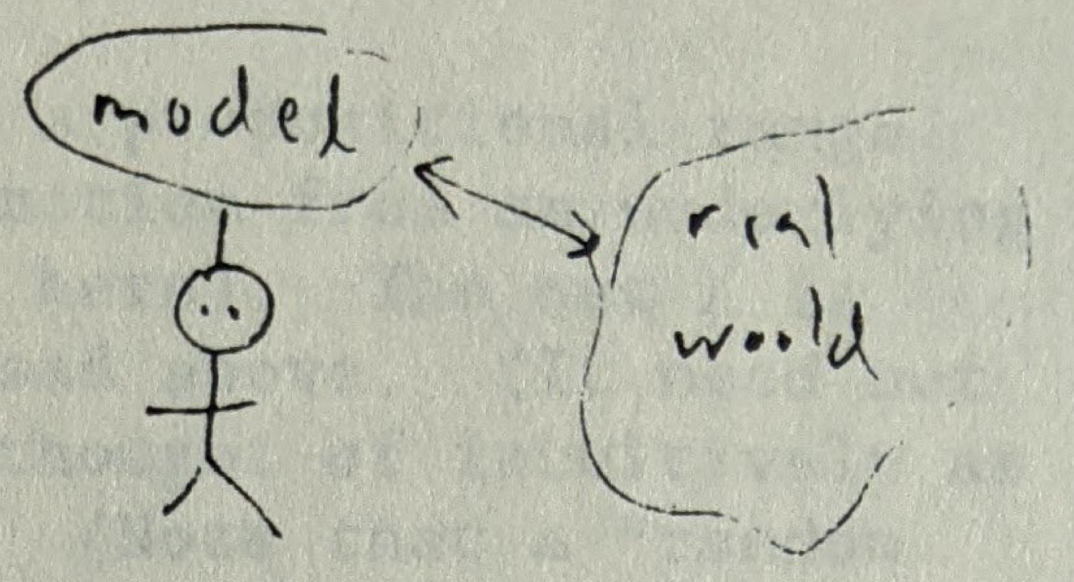

Figure 1

As time goes by one's model changes, partially by assimilating experience, the flow of information, and partially by thinking (internal information-processing) (Figure 2).

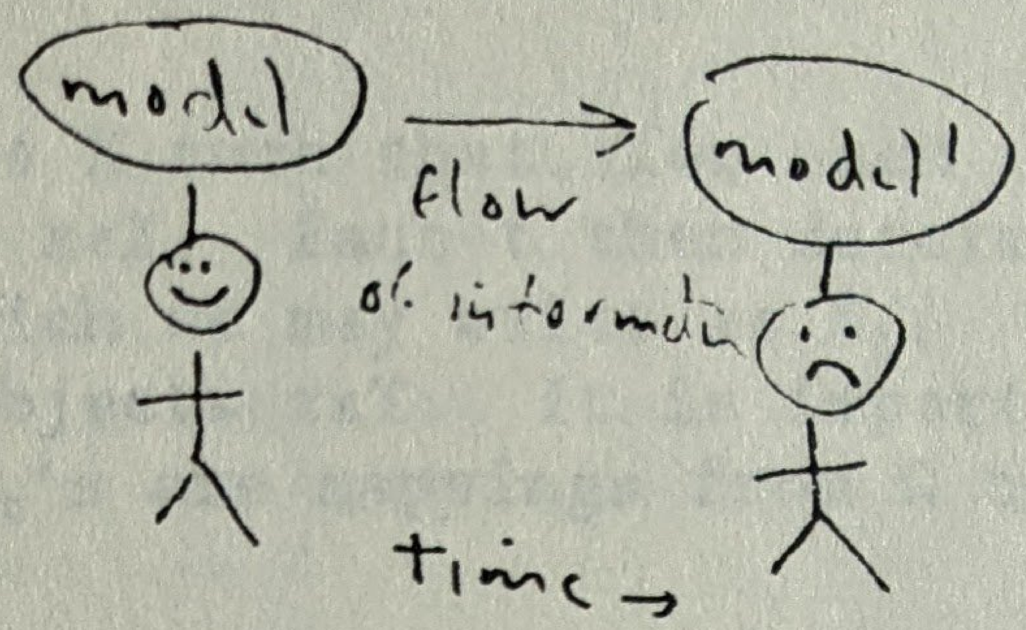

Figure 2

At the same time one is acting (exerting effects upon the world). The quality of one's acts reflects the quality of one's model, and the value of that model is given by the expected value of the results flowing from one's acts. (The capital-theoretic aspects of model evaluation should be noted: The same information can be used over and over, and models can be used as inputs for the production of better models).

The value of a model depends on its degree of correctness (resemblance to the real world), its scope or degree of completeness, and its degree of precision.

III. The Structure of Models

The following material should be in the first chapter of any book on econometrics or applied statistics, but it isn't. The terminology in these fields is also poor, almost designed to sow confusion.

A structure is a set of "objects", having properties and relations among themselves. Abstract structures are typically defined implicitly by axioms. (All of mathematics is the study of abstract structures). Concrete structures have objects in the ordinary sense, existing in space and time.

A proposition is a statement about some structure, which may be true or false (e.g. "object t has property x"). Given propositions P, Q, we can form new propositions: "not P", "P and Q", "P or Q". (Technically, the set of all propositions about some structure is a Boolean algebra).

Let X be a set of properties such that for each object t in a structure, the proposition "t has property x", $x \in X$, is true for exactly one x. (e.g. let t be a coin toss, and X the set {falls heads, falls tails}). This set of propositions is an example of a propositional range, i.e., a set of propositions such that exactly one of them must be true. (The number of propositions in a range may be two, as in coin-tossing, or any greater number, finite or infinite). A similar construction works for X a set of relations.

A random variable is the abstract representation of a propositional range. Technically, a random variable is a "measurable" function from an underlying space Ω to a set X. (I won't discuss measurability here). The set X is exactly the set of properties (or relations) discussed above. (It need not have the structure of the real numbers). Ω may be thought of intuitively as the space of possible structures, models or worlds. (Note that a "random variable" is neither random nor a variable, an example of the dreadful terminology mentioned above.)

Consider again a structure and a set of properties X such that, for each object t, "t has property x" is true for just one $x \in X$. Each t then determines a propositional range, hence a random variable which we may write as X_t. Thus we have a family of random variables indexed by objects $t \in T$. It is important that many r.v.'s come in families. (All of the X_t's are mappings from Ω to X).

The set X itself is what statisticians would call a sample space: repeated observations on objects t yield in general different points in X. (If probabilists use the term, they usually apply it to Ω, not X, another example of terminological confusion).

A stochastic process is an indexed family of random variables. Usually the index is interpreted as time, but it need not be. We can index by the objects of any structure: places, people, countries, etc. (The index is usually called the parameter of the process, which invites confusion with the very different concept used by statisticians).

A parameter is a random variable that is not one of an indexed family, but stands alone. (Example: repeated tossing of a coin with unknown bias θ; let X_t be the outcome of toss t. (t=1, 2, ... Conditional on θ, the X_t's are iid with $Pr(X_t = head|\theta) = \theta$. The r.v.'s are the X_t family, together with θ.

One familiar with ordinary statistics might react to this definition as follows: a parameter is an unknown constant, not a random variable. This again shows confusion, this time deriving from the so-called "frequentist" concept of probability. Yes, a parameter is an unknown constant, but so is every random variable -- e.g. X_3 is either "head" or "tail", we (usually) don't know which.

Finally, lets take a closer look at Ω, the set of "all possible worlds". Each "possible world" $\omega \in \Omega$ specifies the true value of every r.v. (including all parameters). One may simply identify ω with this specification mapping from r.v.'s to their true values. Ω is then identical to the cartesian product of the sample spaces of all random variables. (Example: in a stochastic process where the index set T is time, Ω is simply the set of all possible time series.)

In probability theory this is often done. But in econometrics and other applied areas it is not a good idea, for the following reason. The model in one's mind, and its associated set of r.v.'s, represents only a small fragment of the world. One should hold open the possibility of adding new properties, new relations, and new objects to one's model. The importance of this point cannot be overemphasized, and will play a crucial role in the discussion of causality below. (Incidentally, we may also wish to drop some objects from our models if, e.g., they turn out to be fictitious -- cf. the planet Vulcan in astronomy or the study of demonology).

IV. Probability

Finally, a way of expressing uncertainty is needed. The most common -- but not the only -- way of doing this is by probability. Probability assigns numbers to proposition, Pr(S) being most naturally interpreted as the "degree of belief" that S is true, Pr(S) = 1 being complete certainty that S is true, Pr(S) = 0 being certainty that S is false. More generally, $Pr(S|T)$ is the degree of belief that S is true, contingent on the assumption that T is true ("conditional" probability). "Degree of belief" can be made operational in terms of hypothetical betting behavior (Ramsey, deFinetti, et. al.).

A probability distribution assigns probabilities to all the members of a propositional range -- or, equivalently, to the random variable that represents it. The natural abode of probability is on Ω, the space of "all possible worlds." This not only determines the distributions of all random variables (by projection), but also their joint distributions, and thus dependence-independence relations among them. (Technically, probability is a measure on the field of measurable subsets of Ω, with $Pr(\Omega) = 1$).

Many questions arise. Why should probability be the preferred way of representing uncertainty? Is probability objective or subjective? What relation holds between probability and frequency?

The value of models derives from their aid in making decisions (section II above). Alternative modes of representing and processing information are to be evaluated by this (and the move will tend to supplant the less efficient by the general principle of natural selection.) Now, a somewhat persuasive body of argument has arisen (Ramsey, deFinetti, Jeffreys, Lindley, Savage, Berger, DeGroot, et. al.) which concludes that a necessary condition for "coherent" decisions is that one be a Bayesian. (I won't discuss coherence further except to say that its lack is analogous to having an inconsistent "cyclical" preference order).

Being a Bayesian has three facets:

(i) representation rule: Uncertainty at any one time is to be represented by a probability distribution over all random variables (including parameters).

(ii) <u>updating rule</u>: Having observed that proposition F is true, update your probability distribution by conditioning on F. That is, for any proposition S, the new "posterior" Pr(S) equals the old "prior" Pr(S|F).

(iii) <u>action rule</u>: Act so as to maximize expected utility, expectation being with respect to your current probability distribution.

Is probability objective? "Objectivity" has several meanings. The first is, representing what is "really out there." In this sense, the judgment Pr(S) = p, where $p \neq 0,1$, is not objective, but a confession of ignorance: S is true or false, but we don't know which. Pr(S) charges over time, and may eventually hit 0 or 1, at which point we attain certainty.

But "objectivity" can also mean: judging in conformity to the evidence available. Different models may arise largely because different people have different streams of experience, your probabilities reflecting the perspective from which you view the world which arises from your particular experiences. (Indeed, this "perspectival" view of probability seems to me to be the most adequate interpretation of what probability "is". There is a school of what might be called "objectivist Bayesians" -- Keynes, Carnap, Jeffreys, Jaynes -- who have tried to devise principles to select the "grand prior" to start the process, but these attempts in my view are totally inadequate).

A word about the "frequentist" interpretation of probability, the one apparently favored by most econometricians, though few have thought very hard about it. Consider a stochastic process X_t, t = 1, 2, ..., t being time, the X_t's having a common sample space X (e.g. X = {heads, tails} for coin tossing). Instead of attaching probabilities to propositions, frequentists attach them to "events" in the sample space itself, e.g. Pr(heads), this being interpreted as the limiting ($t \to \infty$) long-run relative frequency of heads among the X_t's. (Thus the "probabilities" are unknown, and must be estimated by past relative frequencies). Certain problems arise: What if the index t is not time? What if the coin melts before t = ∞? How do we know there is a limit? Even if it exists, why should it be of interest for decisions? These difficulties rule out frequentism as a general ground for probability, but there is one important case where it makes sense: When t is time and the X_t process is <u>stationary</u>. In this case, relative frequencies will converge with probability one (by Birkhoff's ergodic theorem), and this limit may be thought of as the "true" probability. (But this does not mean that Bayesian procedures should be abandoned for incoherent frequentist methods even in the stationary case: (i) Bayesian updating moves Pr(X_t is heads) toward the "true" probability (ii) in general, conditioning destroys stationarity in the short-run future, e.g., in the Markov chain with the transition matrix of Figure 3, the long-run frequency of heads is .5, but Pr(X_t = head|X_{t-1} = head) = .99, not .5).

X_{t-1} \ X_t	head	tail
head	.99	.01
tail	.01	.99

Figur. 3

In summary, "frequentist" thinking has the effect of squeezing probability models into a stationary mould, and of sowing confusion, not least the confusion between frequencies and probabilities.

V. Causality

Causality -- the sense that things exert influence on each other -- has deeper intuitive roots than probability. Every sentence in the English language has a verb, and most verbs appear to express causality, in particular almost all transitive verbs -- carries, constructs, eats, plants, etc.

Causality underlies economic theory. Production functions express causal relations from inputs to outputs (and marginal product isolates the influence of a particular input). The flow of information, the formation of prices, the negotiation of contracts, the internal workings of organizations are other examples.

As noted above, the value of models arises from their aid in selecting actions. But to act is to exert influence upon the world. In this sense models themselves enter into causal relations "externally", via the minds they inhabit. Our aim here, however, is to examine the role of causality in the internal structure of models.

The causality concept plays a somewhat furtive role in econometrics. There are occasional references to papers by Simon and Wold. The best known use is "Granger causality" (the concept is due originally to Norbert Wiener), which unfortunately is tied to a conventional significance test, and is thus incoherent according to Bayesian lights (see Zellner's critique).

But this apparently limited role for causality is misleading. I believe that all of the central issues in econometrics tie in with this concept -- e.g. the specification problem, which tries to find the "correct" model (what is a correct model anyway?); the issues surrounding simultaneous equations, such as structural equations vs. reduced forms, least-squares bias, etc.; the problem of errors in variables; the Lucas critique; the concept of exogeneity; the role of parameters. Clarifying the relationship between causality and probability is the key to all of these.

When all this gets worked out, it will show that not merely the content of econometric models, but the form of econometric procedures themselves, will derive from economic theory. (The same applies to all of statistics, for that matter).

Oddly enough, other social sciences have made greater explicit use of the causality concept than has econometrics (see Blalock). Specifically, they use the path analysis of Sewell Wright, which comes from genetics. (The results are not too prepossessing).

Why bother with causality? The answer is the same as that given above for probability -- namely, models constructed causally tend to be better than those that are not, in the sense of leading to more effective actions. Indeed, the argument here is easier to grasp than the "coherence" argument for probability. But first we need a better fix on what causality is, and how to represent it.

There is some interesting recent work on causality by statisticians and computer scientists -- Lauritsen, Darroch, Speed, Pearl, et. al. See
Whittaker
Pearl
Lauritsen

In this work, causality is represented as an <u>acyclic directed graph</u>. A directed graph consists of a set of "nodes" and a set of ordered pairs of these nodes. If (a,b) is one of these ordered pairs, represent this by an arrow going from a to b. Example: Figure 4 represents nodes {a,b,c,d}, pairs (a,b), (a,c): An acyclic graph is one in which there is no directed chain of arrows going in a circle, as in Figure 5.

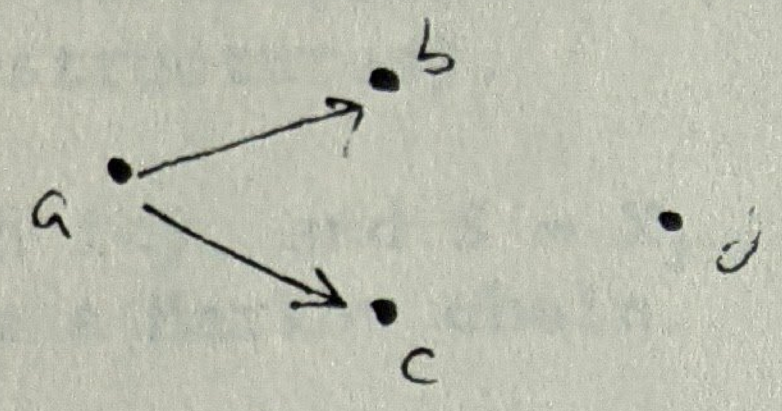

Figure 4

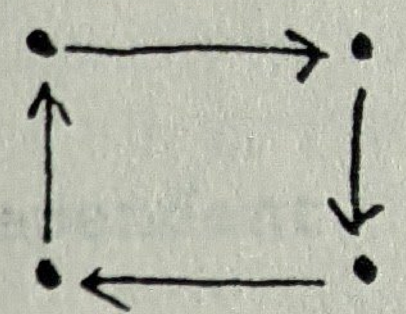

Figure 5

In the work cited above, the nodes are interpreted as random variables, and an arrow (a,b) means, intuitively, that "causal influence" flows from a to b. This vague concept is pinned down by the corresponding joint probability distribution. First, let $\rightsquigarrow$ (wavy arrow) be the transitive closure of $\rightarrow$ (that is, $a \rightsquigarrow b$ if there is a chain of arrows $a \rightarrow c \rightarrow d \rightarrow \ldots \rightarrow b$ beginning with a and ending with b. $a \rightsquigarrow b$ means there is an <u>indirect</u> causal influence from a to b). Now, given random variable X, let S be the set of random variables Y such that $Y \rightsquigarrow X$, and let T be the set of r.v.'s $Y \neq X$ such that $X \rightsquigarrow Y$ is <u>false</u>. (See Figure 6. Note that S is a subset of T, by acyclicity).

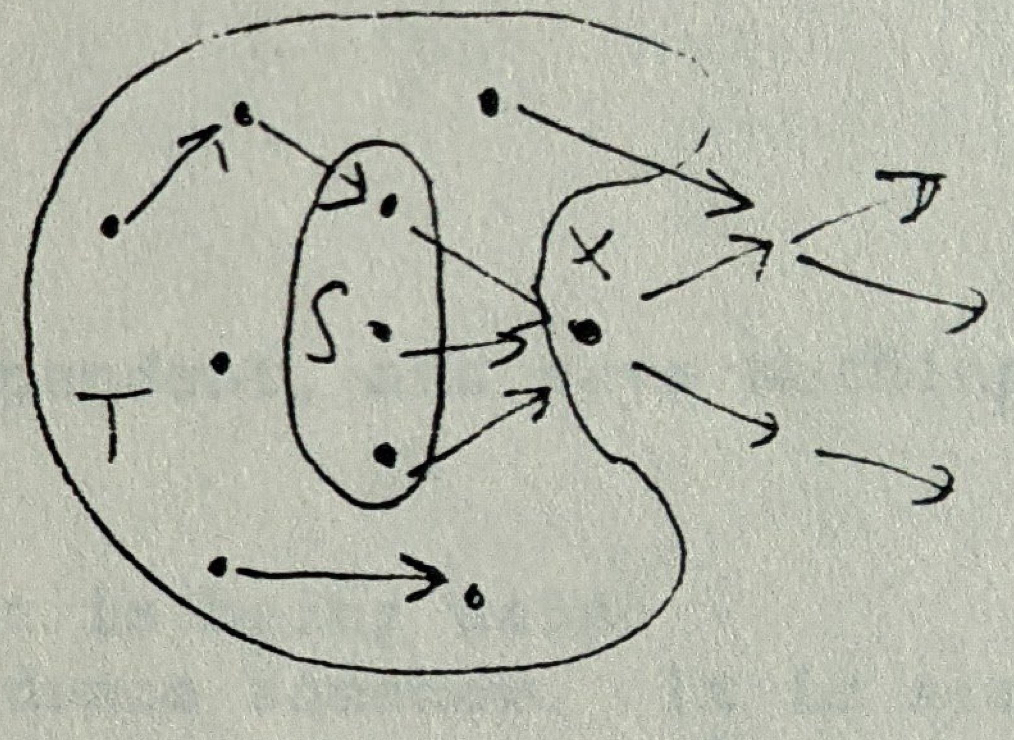

Figure 6

(1) Then $Pr(X|T) = Pr(X|S)$.

The basic idea in this equation is <u>conditional independence</u>: Given the information in random variables S, the additional information in the remaining r.v.'s T\S ("T but not S") does not help in determining X. A more symmetric way of writing this same property is Figure 7: Given S, X and T/S are independent (S separates, or screens off, these from each other).

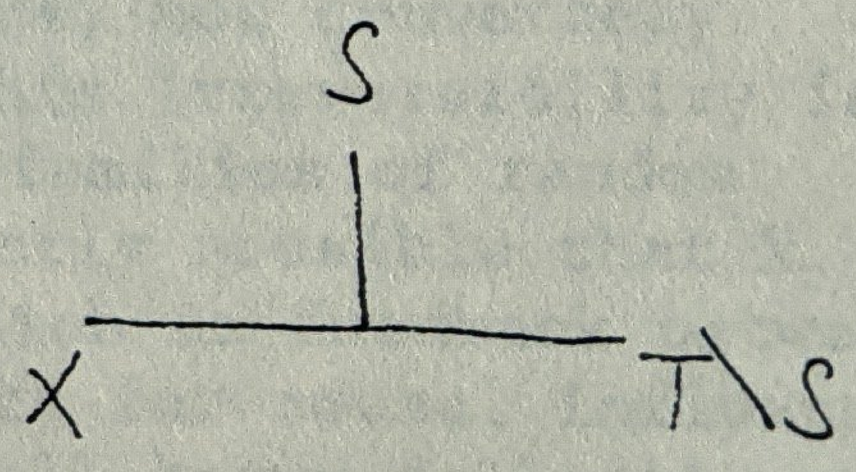

Figure 7

Here are some examples.

(i) No arrows at all. Then all variables are independent. (This is easy to see: for any X, S is empty and T is all other r.v.'s.)

Note that, in general, the fewer arrows the stronger the statement being made.

(ii) Let $X_1, .. X_n$ be the r.v.'s and let $X_i \rightarrow X_j$ if and only if $i<j$. Then, for any X_j, $S = T =$ all X_i with $i<j$. Thus (1) or Figure 7 say nothing at all, and this "causal structure" is compatible with <u>any</u> joint distribution. (Since the ordering of the X's is arbitrary, there are n! such structures).

(iii) $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n$. For any X_j, $T =$ all X_i with $i<j$, and $S = X_{j-1}$. Then (1) is precisely the Markov property: $X_1, \ldots X_n$ form a Markov chain.

(iv) Reverse all arrows in (iii). Then $X_n, \ldots, X_1$ is a Markov chain. But this is the same class of distributions as in (iii), since a time-reversed Markov chain is still Markov!

As a final example, given r.v.'s X,Y,Z, suppose X,Y are dependent, but are independent conditional on Z:
This may be represented by

$X \rightarrow Z \rightarrow Y$ or by $Y \rightarrow Z \rightarrow X$

or by $\begin{tikzcd}[row sep=tiny] & X \\ Z \arrow[ur] \arrow[dr] & \\ & Y \end{tikzcd}$, but <u>not</u> by $\begin{tikzcd}[row sep=tiny] X \arrow[dr] & \\ & Z \\ Y \arrow[ur] & \end{tikzcd}$

(Exercise: Show this last graph implies X,Y are independent, and says nothing about conditional independence).

What is one to make of all this? Something important is being said: Causality certainty ties up with conditional independence somehow. It is also clear that, in general, the distribution does not determine the causal structure ("correlation is not causation") nor is it clear yet that knowing the causal structure is of any help in making decisions.

What needs to be done is to go back to the intuitive roots of the causal concept (as noted above these are much stronger than the intuitions underlying probability).

First, causal influence goes from the past to the future, not conversely. This explains acyclicity, since time does not loop. This irreversibility is compatible with "feedback", however: let X_t, Y_t be two families of random variables indexed by time, and let $r<s<t$. It is perfectly possible that X_r causally influences Y_s, which in turn influences X_t, which is feedback between X and Y. Finally, note that while time lag is necessary for causal influence, it is not sufficient (Post hoc ergo propter hoc is a well-known fallacy).

Second, it is not quite correct to speak of causal influence flowing from one random variable to another. Rather, it flows from a real object at a given time to itself or another object at a later time. Recall that random variables are often indexed by objects. To say that r.v. X indexed by i influences Y indexed by j is an oblique way of referring to the influence of i upon j (which may manifest itself by probabilistic dependence between X and Y).

I have just begun to work out the implications of this second observation, and for the most part I shall "go with the crowd" and discuss causation among r.v.'s. (The kind of causal influence emanating from i to j is itself a random variable Z indexed by the pair (i,j), and having a probability distribution jointly with the other r.v.'s.)

In contrast to the treatment above, which takes causality as an adjunct to conditional independence, we propose to put causality first, and derive conditional independence from it. (The excellent book by J. Pearl ______________ takes a similar approach.)

Example: Let X = barometer reading today, Y = weather tomorrow. These r.v.'s are, we may assume, dependent, but clearly there is no causal influence from one to the other (even though X precedes Y in time). Instead there is a third variable that causally influences both of them: Z = atmospheric conditions today ($X \leftarrow Z \rightarrow Y$). Thus, given Z, X and Y are independent, and the "spurious" correlation between X and Y arises from their common cause. There are any number of patterns of this sort -- e.g. X = being in the hospital today, Y = dying tomorrow. The positive dependence between these may be due to Z = being sick yesterday (more controversial: X = smoking, Y = getting lung cancer. Is there a causal flow from X to Y, or do they have a common cause Z = genetic predisposition, as R.A. Fisher thought?)

The role of <u>parameters</u>. As a rule, parameters causally precede the other r.v.'s in a model. Example: coin tossing with unknown bias θ. Given θ, the X_t's are independent, namely, $\Pr(X_t = \text{head}|\theta) = \theta$. Thus Figure 8 is the causal structure. (Think of the parameters as being fixed before the model unrolls).

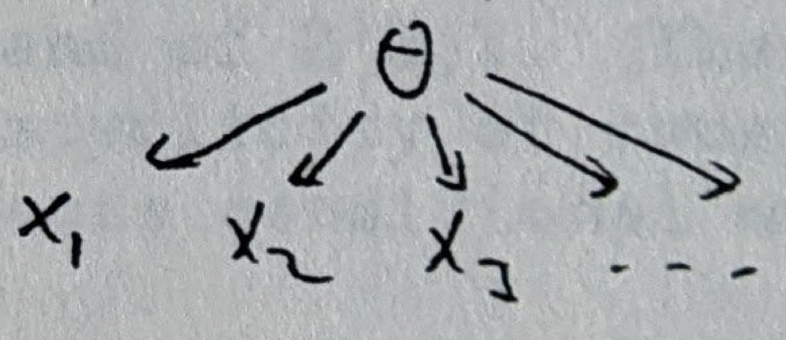

Figure 8

One essential point. Finding the causal structure usually means adding "explanatory" random variables to an existing model, even unobservable ones. In this sense only does it make sense to ask whether a model has the "correct" random variables ("correct" meaning <u>causally closed</u> -- i.e., a model having a true causal flow structure that is compatible with the joint probability distribution).

<u>Causality and Probability</u>

It is a striking fact that Bayesians, who have the most adequate conception of what probability is, hardly ever discuss causality (this is true even by the natural scientists among them, such as Jaynes and Jeffreys). The reason is, that a complete probability distribution over "all possible worlds" is all one needs to know or can know. There is no room left for additional information provided by causation or the like.

Granted this, there is still a problem. Who has such a comprehensive distribution? The answer is, no one. We don't run around with distribution in our heads. Even assessing the distribution of a single real-valued random variable is not trivial. To extend this to a joint distribution over an indefinitely large number of random variables is a formidable task.

Once we view distributions as <u>achievements</u> rather than givens -- due to the limitations of the human mind -- things begin to fall into place. First, a role for causality arises, and a fundamental one at that. (Second, a new basis for all of statistical inference arises, based on the reduction of complexity costs. This is the so-called <u>post-Bayesian</u> approach, and will be discussed briefly below).

Thesis: Causality is indispensable for assessing joint distribution over many random variables. Let X = spring weather, Y = fall crop size. Then we may suppose causal influence flows from X to Y. The joint distribution $Pr(X,Y)$ merely shows dependencies. Now this can be built up in two different ways as $Pr(X)\ Pr(Y|X)$ and as $Pr(Y)\ Pr(X|Y)$. I suggest that it is easier and more natural to build things up in the former way, that is, to assess effects conditional in causes rather than the other way around. The reason is that it is more useful to know the effects on the future of what happens now (in particular, of what we do now) than the reverse, so that the human mind is attuned to this direction. (Note in this connection that the use of Boyes theorem, which goes from $Pr(Y|X)$ to PrLX|Y), used to be known as "inverse probability", or the "probability of causes". As Kicrkegaard said, we live forward but reason backward).

Another consideration makes this tendency self-reinforcing. The joint distribution of X,Y,Z, say, can be built up in 3! = 6 ways, each of these ways having 3 blocks, say the probabilities of X, of Y|X, and of Z|X,Y. These "fit together", while, say, Y|X and X|Y do not. Now, the acyclicity of causal influence ensures that, if we consistently assess effects conditional on causes, then the pieces will indeed fit together.

Now consider a model having random variables X, and let Y be a different set of random variables. In general, the causal structure of the X's depends on the Y's having "normal" values. (Just as a production function depends on the presence of gravity, normal temperatures and an atmosphere, none of which are explicitly mentioned). Under "abnormal" circumstances, some of the causal connections may be switched on or off.

Consider a model indexed by time t and another index i (the objects in the model). The model has a <u>stationary causal structure</u> if the following holds: if causal influence flows from X_{is} to X_{jt} ($s<t$), then it flows from $X_{i,s+\theta}$ to $X_{j,t+\theta}$ for all θ. This concept expresses the "invariance of the laws of nature" and is more plausible than the stronger assumption of stationarity per se (see Figure 9). (More generally, some of those

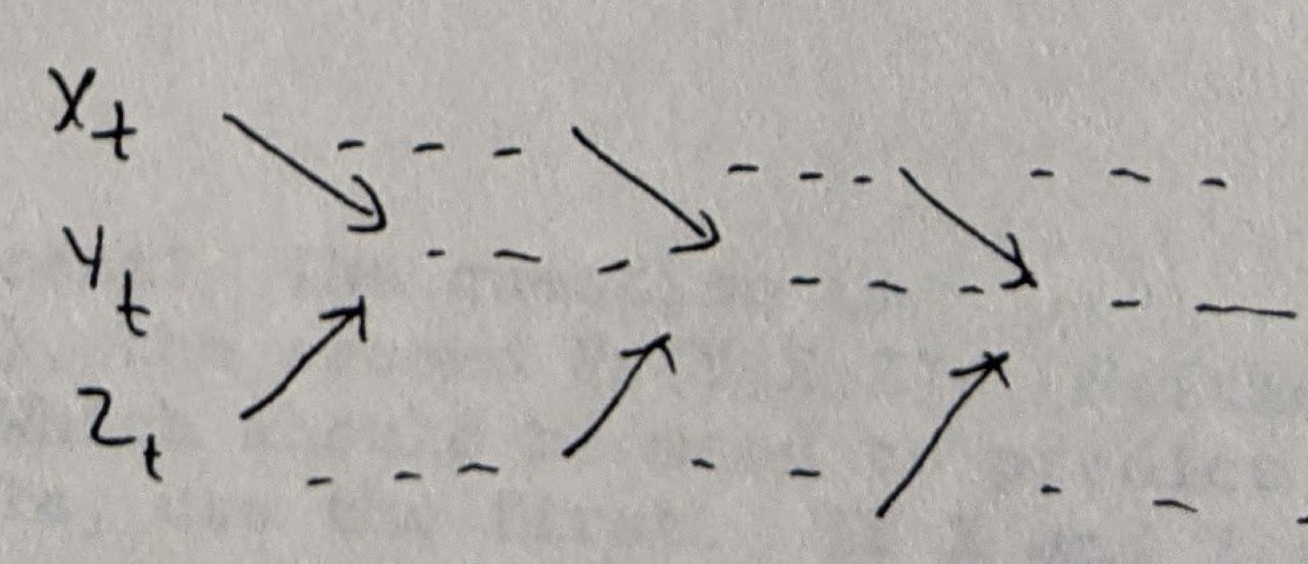

Figure 9

causal connections may be switched on or off at various times, perhaps in response to outside influence). For example, the configuration of the solar system never repeats, but its laws of evolution are constant.

Causality and Econometrics

We end with some thoughts of how all this relates to econometrics.

First, an econometric model is a probability distribution. It may be represented by a set of equations, but these are precisely that, a representation. It follows that two different systems of equations yielding the same probability distribution are picturing the same model.

To be more exact, models do not usually give the full joint distribution over all random variables considered, but the conditional distribution of some of them with respect to the others -- say $Pr(Y|X, \theta)$. Here Y are usually called the endogenous variables, X the exogenous and θ the parameters. Further, if the X's and Y's are time indexed, this joint Y-distribution may be factored as

$$Pr(Y_1|X,\theta)\ Pr(Y_2|Y_1X,\theta)\ldots Pr(Y_T|Y_1\ldots Y_{T-1}\ X\ \theta),$$

the conditioning Y's then being called "predetermined."

Let us first adopt a purely probabilistic view of all this, ignoring causality entirely (so that "endogenous" and "exogenous" are just names for the variables we choose to place on the left or right side of the conditioning bar). From this point of view, the question of whether the model is "correct" or "incorrect" is meaningless: any collection of random variables has a joint distribution. To be specific, suppose in the usual manner you "regress Y on X". This is an estimation procedure for the conditional probability $Pr(Y|X)$.

Suppose X = Spring weather, Y = Fall crop size. Then regressing Y on X is routine. Is it legitimate to regress X on Y? The answer is yes. (Assuming some conditions that justify regressing -- e.g. (X_t,Y_t) are joint normal and iid given the means and covariances). This estimates $Pr(X|Y)$. Certainly Y provides information about X, even though there is no flow of influence from Y to X. (Any inference about the past is of this form.)

Now consider Y and possible "explanatory" variables X,Z (X,Y and Z may be vectors). Regressing Y on X yields, say,

(2) $\quad Y = aX + \text{residual}.$

Regressing Y on X and Z yields, say,

(3) $\quad Y = bX + cZ + \text{residual},$

where, in general $a \neq b$. Which model is correct? The question is meaningless. The first is estimating $Pr(Y|X)$, the second $Pr(Y|X,Z)$. Neither of these is "more correct" than the other. Which should be used to predict (or estimate) Y? Well, if only X is available, use the first. If X and Z are both available, use the second. (Why not then regress Y on everything in sight? The answer involves a Bayesian critique of the standard regression model which I won't discuss here.)

Thus in a model where causality is ignored life is simple. There are no intrinsically exogenous or endogenous r.v.'s, no "incorrect" models, just a joint probability distribution. (Though causal considerations would have been needed to assess the distribution in the first place).

When is it satisfactory to ignore causality and focus exclusively on probability like this? Roughly, the answer is: when one is concerned only with causal influences <u>from</u> the model <u>to</u> the rest of the world. (Example: the "state of nature" in decision theory, which influences consequences, but cannot be influenced by them). But to assess influence <u>from</u> the rest of the world <u>to</u> the model, the internal causal structure of the model is needed.

As an example, consider the weather. Insofar as the weather influences, but cannot be influenced by, human activities, it suffices to know its probability distribution. But if human activities influence the weather (e.g. cloud-seeding, PCB's), the causal structure of the weather system is needed.

Consider a model involving X,Y and Z again, and suppose that X, but no other variable can be influenced directly from outside. Which of the regressions above correctly assesses the indirect effect on Y? The answer is, it could be the first, the second, or neither. Consider various causal structures.

In Figure 10, X and Z are independent. Then a=b in (2) and (3), and both regressions are correct.

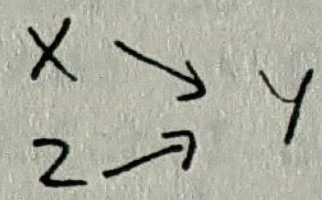

Figure 10

In Figure 11, only (2) is correct. We don't want to "hold Z constant" since the influence of X is transmitted through Z (In this case b=0 in (3)).

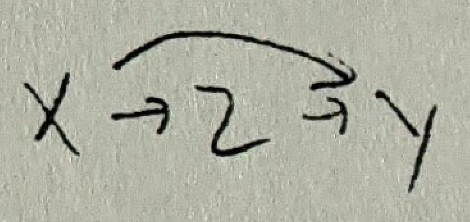

Figure 11

In Figure 12, only (3) is correct. Z causes a "spurious correlation" between X and Y which must be netted out.

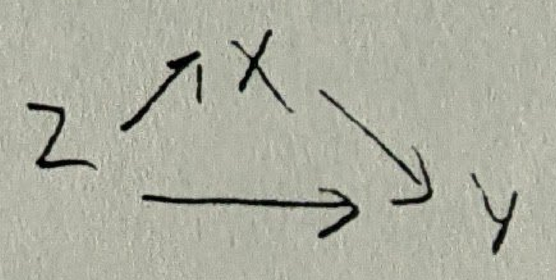

Figure 12

In Figure 13, neither model is correct. X has no influence on Y, yet in general, neither coefficient a nor b will be zero.

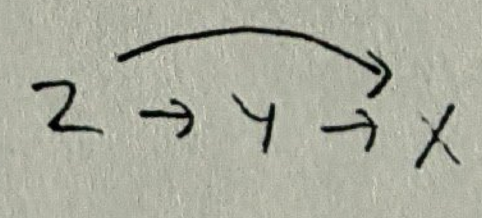

Figure 13

This is a good place to discuss "Granger causality". Let X_t, Y_t be two families of random variables indexed by time. Suppose there is no causal influence from Y to X (there may be some from X to Y). According to Granger, one tests this by the consequence that, for any t, X_t should be independent of all past Y's given all past X's. But this is a fallacy. What prevents there from being a third family Z_t which influences both the X's and Y's? (cf. Figure 12). If so, past Y provides information about past Z, and thus indirectly about present X_t beyond what past X provides.

A word about Sims' "atheoretical" approach. In effect, this says that, for any r.v.'s X and Y, if X is earlier than Y, then X exerts causal influence on Y. This is the "post hoc ergo propter hoc" fallacy.

A word about simultaneous equations. This is not a good setup for capturing causal structure, since it violates the acyclicity condition. Early critics Strotz and Wold tried to remedy this by imposing an ordering on variables within each period (conditional causal chains). The trouble here was that, in terms of their economic interpretations, the variables overlapped in time and therefore could not have had the causal structure indicated.

Most econometric models are expressed by regression equations. If these are interpreted as expressing causal structure (influence flows from the independent to the dependent variables), then the question of correctness does become germane. Which econometric models are then correct in the causal sense? Probably none. This conclusion arises from several considerations. First, the nature of the data that the r.v.'s represent: it is noisy and aggregated in space and time. Second, one can almost always think of omitted relevant variables: the structures are not causally closed. Third, almost every model involves arbitrary decisions concerning functional form, lag structure, etc.

Thus all our models are at best approximations to the "true" causal structure. This is OK: good approximations may work well. But this casts grave doubt over significance testing in general. Why ask a question to which the answer is known in advance?: all null hypotheses are false. And in fact there is an alternative approach to all of statistical inference which I call the <u>post-Bayesian</u> approach, which is based on complexity costs and the limited capacity of the human mind discussed above.