

FUNDAMENTALS OF STATISTICAL CAUSALITY

© A. P. Dawid 2007

RSS/EPSRC Graduate Training Programme

University of Sheffield
3–7 September 2007

Version of September 17, 2007

Preface

Traditionally, Statistics has been concerned with uncovering and describing associations, and statisticians have been wary of causal interpretations of their findings. But users of Statistics have rarely had such qualms. For otherwise what is it all for?

The enterprise of “Statistical Causality” has developed to take such concerns seriously. It has led to the introduction of a variety of formal methods for framing and understanding causal questions, and specific techniques for collecting and analysing data to shed light on these.

This course presents an overview of the panoply of concepts, with associated mathematical frameworks and analytic methods, that have been introduced in the course of attempts to extend statistical inference beyond its traditional focus on association, and into the realm of causal connexion. Emphasis is placed on understanding the nature of the problems addressed, on the interplay between the concepts and the mathematics, and on the relationships and differences between the various formalisms. In particular, we show how a variety of problems concerned with assessing the “effects of causes” can be fruitfully formulated and solved using statistical decision theory.

Our emphasis is almost entirely on problems of “identification”, where we suppose the probabilistic structure of the processes generating our data is fully known, and ask whether, when and how that knowledge can be used to address the causal questions of interest. In the causal context, the more traditional statistical concern with estimation of unknown probabilistic structure from limited data is a secondary (admittedly extremely important, and currently highly active) enterprise, that will hardly be addressed here.

For reasons of space and coherence, our emphasis is also largely restricted to understanding and identifying the effects of applied causes. The problem of identifying the causes of observed effects raises many further subtle issues, both philosophical and mathematical, and would take us too far afield.

Acknowledgments

Vanessa Didelez and Sara Geneletti have been major collaborators in the original work reported here. I am grateful to Antonio Forcina for helpful comments.

Contents

I	GENERAL CONSIDERATIONS	9
1	What's So Hard About It?	11
1.1	Experimental studies	11
1.2	Observational studies	11
1.3	Some examples	12
1.4	Difficulties	14
1.4.1	Common cause	14
1.4.2	Complete confounding	14
1.4.3	Reverse/two-way causality	14
1.4.4	Selection	14
1.4.5	Regression to the mean	14
1.4.6	Simpson	15
1.4.7	Promotion and prevention	15
1.5	Randomization	15
2	Causal Questions	17
2.1	Effects of Causes and Causes of Effects	17
2.2	Similarities and differences	17
2.3	Hypotheticals and counterfactuals	18
2.4	Probability and Decision Theory	18
2.5	Potential responses	19
2.6	What can we know?	20
3	Formal Frameworks	21
3.1	Stochastic model	21
3.2	Potential response model	21
3.3	Structural model	22
3.3.1	Extended structural model	22
3.4	Functional model	23
3.5	Fitness for purpose	23
4	Some Assumptions	25
4.1	Exchangeability	25
4.1.1	Simple case	25
4.1.2	de Finetti's theorem	25
4.1.3	Conditional exchangeability	26
4.1.4	Causal inference	26
4.1.5	Extensions	26
4.1.6	PR approach	27
4.2	Treatment-unit additivity	27
4.3	Stable unit-treatment value assumption	28
4.4	Is there a "fundamental problem of causal inference"?	28

II	ALGEBRAIC AND GRAPHICAL REPRESENTATIONS	31
5	Conditional Independence	33
5.1	Properties and axioms	33
5.2	Further axioms?	35
5.3	Extension to non-stochastic variables	35
5.4	Conditional independence as a language for causality	35
6	Directed Acyclic Graphs	37
6.1	DAG representation of a distribution	37
6.2	Factorization of the density	38
6.2.1	Ancestral sets	38
6.3	Conditional independence properties implied by a DAG	38
6.4	Moralization	39
6.5	d -separation	41
6.6	Markov equivalence	43
6.7	Influence diagrams	44
7	Causal Interpretations Of DAGs	47
7.1	Intervention DAGs	47
7.1.1	Seeing and doing	49
7.2	Augmented DAGs	49
7.2.1	Extensions	51
7.3	Functional DAGs	52
7.3.1	Latent variable models	53
7.4	Functional intervention models	53
III	SPECIAL TOPICS	55
8	Computing Causal Effects	57
8.1	General approach	58
8.2	DAG models and Pearl's " <i>do</i> calculus"	59
8.2.1	Back-door and front-door	60
8.3	Nonidentifiability of causal effect	60
8.3.1	Bow-pattern	60
8.3.2	Parent-child bow	61
9	Confounding And Sufficient Covariates	63
9.1	Example: Normal regression	63
9.2	No confounding	64
9.2.1	Potential responses	64
9.3	Confounding	64
9.4	Sufficient covariate	65
9.5	Allocation process	66
9.6	Potential responses	67
9.7	Confounding	67
9.8	Nonconfounding	67
9.8.1	Other conditions	69
9.8.2	'No unobserved confounders'	69
9.9	Deconfounding	69
9.9.1	Complete confounding	69
9.9.2	External standardization	70
9.10	Average causal effect	70

9.10.1	Potential responses	70
10	Reduction Of Sufficient Covariate	71
10.1	Reduction of effect on Y	71
10.1.1	Minimal response-sufficiency	72
10.2	Reduction of effect on T	73
10.2.1	Allocation process	74
10.2.2	Minimal treatment-sufficiency	75
10.3	Propensity scoring in practice	75
10.4	A complication	76
10.5	Joint reduction?	77
11	Instrumental Variables	79
11.1	Causal inference	81
11.2	Null hypothesis	81
11.3	Linear model	81
11.4	Binary case	82
11.4.1	Instrumental inequalities	82
11.4.2	Causal inequalities	82
11.4.3	General discrete variables	84
12	Effect Of Treatment On The Treated	85
12.1	Special cases	85
12.1.1	Allocation variable	85
12.1.2	Potential responses	85
12.2	Uniqueness of ETT	85
12.3	Identifying ETT	86
13	Dynamic Treatment Strategies	87
13.1	More structure	89
13.2	Other conditions	89
13.3	PR approach	90

Part I

GENERAL CONSIDERATIONS

Chapter 1

What’s So Hard About It?

Causal inference is surely the most basic and straightforward of statistical problems. Following Fisher, almost every introductory Statistics course will illustrate the two-sample t -test using examples such as an experiment where two different fertilisers are randomly assigned to various agricultural field-plots. The population mean responses under the two treatments are the parameters of interest, and the difference between these can be considered as the “causal effect” of choosing to use one rather than the other. This can be estimated (and that estimate hedged about with a measure of its uncertainty) by the difference in sample means for the treated and untreated units. Where’s the problem?

1.1 Experimental studies

A good instructor will emphasize the importance of good statistical *design* as a prerequisite for the above analysis. Before Fisher introduced and insisted on randomization, agricultural researchers would typically lay out their experiments by deterministically assigning the treatments to plots, according to their best judgments, in an attempt to balance out the effects of suspected fertility gradients *etc.* But although there is nothing to stop us applying the mathematics of the two-sample t -test to the results of such an experiment, its validity and relevance can be called into question. Even more problematic cases arise in studies of new medical treatments, where an enthusiastic doctor can produce excellent results that others can not replicate—perhaps because he gave his favoured treatment to carefully cherry-picked patients, perhaps because they received extra special care, or for many other reasons that can undermine the apparent message of the data. These problems gave rise to what is perhaps the most important medical advance of the 20th century: the randomized controlled clinical trial (RCT).

1.2 Observational studies

For all that carefully designed experiments are highly desirable, and should be conducted wherever possible, there are many cases where they can not be performed, for practical or ethical reasons. In this case, we may have to rely on *observational data*, where the putative “causal factors” are not under the control of the investigator. Most of the “findings” reported under screaming 3-inch headlines in the daily press are taken from observational studies.

Very often the desired interpretation is to extract some “causal conclusion” from observational data. This is where it begins to get tricky...

The problem is that we typically won’t have the right data, or data in the right form, to address the questions of interest. We may be interested in comparing what would happen to a patient (or plot of soil) under various treatments that we might apply: that is, we are really concerned with assessing and comparing possible treatment *interventions*, applied to and compared on one and the same subject—having, necessarily, constant characteristics. However in an observational study the

characteristics of the subjects may not be comparable across the different treatment groups, and even if they are may not be comparable with those of new subjects we are interested in treating. The variables measured, and the conditions under which they are produced and recorded, may not be identical with those in the new situations we wish to understand. In the presence of such “confounding”, we will not know whether to ascribe observed differences between the responses in different treatments groups to the treatments themselves, or to other differences between the groups, whose effects would persist even if all units were treated identically.

In such a case, a naïve analysis, treating the treatment groups as if they were comparable, can be dangerously misleading, and other, more subtle, forms of analysis are required. These in turn rely on complex “causal models”, relating, in some appropriate and, ideally, justifiable way, the structure of the process generating the observational data with the quite distinct process relevant to our causal queries. Such models, and their implications for inference, will be the focus of these lectures. But they must not be applied too glibly: in the absence of objective support from good experimental design, there will usually remain much scope for disagreement as to the validity of the causal assumptions made, and consequently of their implications.

1.3 Some examples

Example 1.1 Cannabis ‘raises psychosis risk’.

From BBC NEWS, 27 July 2007.

Cannabis users are 40% more likely than non-users to suffer a psychotic illness such as schizophrenia, say UK experts. The researchers looked at 35 studies on the drug and mental health—but some experts urged caution over the results.

The study found the most frequent users of cannabis have twice the risk of non-users of developing psychotic symptoms, such as hallucinations and delusions. However, the authors said they could not rule out the possibility that people at a higher risk of mental illness were more likely to use the drug. Study author, Professor Glyn Lewis, professor of psychiatric epidemiology, said: “It is possible that the people who use cannabis might have other characteristics that themselves increase risk of psychotic illness.”

Professor Leslie Iverson, from the University of Oxford, said there was no conclusive evidence that cannabis use causes psychotic illness. “Their prediction that 14% of psychotic outcomes in young adults in the UK may be due to cannabis use is not supported by the fact that the incidence of schizophrenia has not shown any significant change in the past 30 years.”

Questions:

- (i). Will your future mental health be affected if you start to use (or stop using) cannabis?
- (ii). Will the nation’s mental health be affected if steps are taken to control cannabis use?

□

Example 1.2 ‘The facts about fuel’.

From Which?, August 2007.

Mr Holloway said that a colleague of his used to drive from London to Leeds and back, using Shell petrol to go up there and BP fuel to drive back. He was convinced the BP petrol gave better fuel economy, but Ray had another explanation: ‘I pointed out that Leeds is at a higher altitude than London: he was going uphill one way and downhill the other!’

□

Example 1.3 Socialization of children. It is common to regard the environment in which children are raised, and the behaviour of those around them, as major influences on their mental health. Bell (1977), Bell and Harper (1977) found that it was just as likely that a child's behaviour was influencing parental behaviour (or the behaviour of its teachers) as it was that the rearing environment had caused the child's behaviour. \square

Example 1.4 Hormone replacement therapy and coronary artery disease (Prentice *et al.* 2005). Observational research on postmenopausal hormone therapy suggested a 40-50% reduction in coronary heart disease incidence among women using these preparations. In contrast, the Women's Health Initiative clinical trial of estrogen plus progestin found an elevated incidence. Even after age adjustment, estrogen-plus-progestin hazard ratio estimates for coronary heart disease, stroke, and venous thromboembolism in the observational study were 39-48% lower than those in the clinical trial. \square

Example 1.5 Vitamin supplements and mortality. Many observational studies have appeared to indicate that antioxidant supplements reduce the risk of disease. In contrast, randomized controlled trials indicate that vitamins A, E and β -carotene may actually increase mortality (Bjelakovic *et al.* 2004). \square

Example 1.6 Calcium channel blockers. In 1995 non-experimental studies suggested an increased risk of myocardial infarction associated with the short-acting calcium channel blocker (CCB) nifedapine. Concern soon spread to the entire class of calcium antagonists. It took almost a decade to obtain RCT evidence, which showed that long-acting nifedapine is safe. \square

Example 1.7 Traffic cameras. In 1992 33 speed and red light cameras were installed in accident-prone locations in West London. The number of fatal accidents, 62 in the preceding three years, went down to 19 in the following three years (Highways Agency 1997). Subsequently traffic cameras were installed all over London. \square

Example 1.8 Simpson's paradox. The following data refer to 800 patients having a certain serious disease. A new treatment was tried on half the group, the remainder having the standard treatment.

	Recovered	Died	Total	Recovery rate
New treatment	200	200	400	50%
Standard treatment	160	240	400	40%

Table 1.1: Overall results

It seems that the new treatment increases the recovery rate by 10 percentage points. However, when these data are broken down according to the patient's sex, the following figures emerge:

	Recovered	Died	Total	Recovery rate
New treatment	180	120	300	60%
Standard treatment	70	30	100	70%

Table 1.2: Male results

	Recovered	Died	Total	Recovery rate
New treatment	20	80	100	20%
Standard treatment	90	220	300	30%

Table 1.3: Female results

It now appears that, for either sex, the new treatment decreases the recovery rate by 10 percentage points. \square

Example 1.9 Underage drinking. Finally, *New Scientist* (11 August 2007) quotes a Press Release from the University of Missouri-Columbia (“Mizzou”):

Mizzou study shows that possessing a fake ID results in more drinking by underage College students.

□

1.4 Difficulties

The above examples exhibit a range of difficulties, which go under the general head of *confounding*—meaning that the observed “signal” in the data need not be a pure effect of the putative cause under study, but may also be due, in whole or in part, to other factors that vary together with that “cause”.

1.4.1 Common cause

In Example 1.1, the problem is how to distinguish between whether cannabis use causes psychosis, or some other, pre-existing, characteristics cause both cannabis use and psychosis (or some combination of the two effects). Both hypotheses would explain the observed association—but they have very different implications for what to do about it. Iverson’s remark gives some support to the common cause hypothesis.

1.4.2 Complete confounding

In Example 1.2, there is complete confounding between the fuel type and the altitude rise/drop of the journey. Either or both may affect fuel economy, and it is impossible to separate their effects.

1.4.3 Reverse/two-way causality

Parents’ behaviour may affect their child’s behaviour, or *vice versa*, or both. With observational data these very different explanations can be hard to tell apart.

1.4.4 Selection

Examples 1.4, 1.5 and 1.6 serve as important warnings of the perils of interpreting observational evidence. The underlying problem is that those choosing or chosen to receive the treatment may well not be typical of the population at large. People who take HRT or supplements tend to be healthier, so that the observational studies may be measuring the influence of other factors than that under investigation. In Example 1.6 the problem is *confounding by indication*: CCBs were used to treat hypertension, and this itself is a risk factor for myocardial infarction.

1.4.5 Regression to the mean

The apparent accident reduction in Example 1.7 is unlikely to be fully explained as a causal phenomenon, again because of selection effects. If the treatment had consisted of emptying a teaspoon of water onto the chosen sites, rather than installing cameras, we would still have expected a reduction—because the high pre-treatment values at the chosen sites were likely to be, at least in part, the results of random, and hence non-recurring, fluctuations (Stone 2004). A better study would compare the reductions in the treated sites with those of similarly selected sites that were untreated.

1.4.6 Simpson

Example 1.8 is a toy example constructed to highlight the way in which confounding (in this case between treatment and sex) can operate to obscure and even reverse the apparent effect of treatment. It is considered in more detail in Example 9.3 below.

1.4.7 Promotion and prevention

Finally, the Mizzou drinking study of Example 1.9, which is presumably a case of common cause, is interesting in that, while we might not expect that issuing an underage student with a fake ID would have much effect on his drinking habits, removing his fake ID from a student who already has one could have a large effect.

1.5 Randomization

The randomised controlled trial (RCT) is generally taken as a “gold-standard” for the assessment of causal effects. In particular, randomization helps guard against many of the difficulties of interpreting observational evidence displayed by the above examples.

- By ensuring that assignment of treatment is entirely unrelated to other variables, randomization can eliminate confounding and the problem of the common cause. When we compare outcomes in different treatment groups, we are “comparing like with like”, and can therefore ascribe any observed differences in effect to the only real difference between the groups: the treatment applied.
- External intervention to apply a treatment ensures that any observed associations between treatment and response can indeed be given a causal interpretation.
- Because application of treatment necessarily precedes measurement of outcome, the direction of causality is clear.

However, it is often not possible, for pragmatic or ethical reasons, to conduct a RCT; or we may want to interpret observational data collected by others. We therefore need to formalize the principles underlying sound causal inference, in a variety of contexts, in a way that will allow us to decide what can or can not be safely inferred, and under what assumptions—and to understand the meaning, import and applicability of those assumptions.

Chapter 2

Causal Questions

Many different varieties of causal question can be identified, but we shall focus on those to which Statistics can make a useful contribution. Questions such as “What is the true cause?”, or “What is the ultimate cause?” of some effect generate powerful philosophical dust-storms, which we shall steer well clear of. Also, as quantitative statisticians, we shall be more concerned with *measuring* the effect of a putative cause, rather than the binary question of whether or not there is a non-zero causal effect at all.

2.1 Effects of Causes and Causes of Effects

Causal questions come in two principal flavours: questions about the effects of applied causes (“EoC”), and questions about the causes of observed effects (“CoE”).¹

Examples are:

Effects of causes : *I have a headache. Will taking aspirin help?*

Causes of effects : *My headache has gone. Is it because I took aspirin?*

Much of classical statistical design and analysis—for example randomized agricultural or medical experiments—has been crafted to address EoC-type questions. CoE-type questions are generally more of intellectual than practical interest—with the notable exception of legal proceedings to determine liability. They are also much more problematic, both philosophically and methodologically. We shall devote most of our attention to EoC.

2.2 Similarities and differences

There are both similarities and differences between the approaches we shall take to the two kinds of question.

Similarities:

Intervention For both EoC and CoE, it is helpful to consider a “cause” as an external *intervention* in a system—for example, the decision to take aspirin—rather than an outcome chosen by Nature.

Contrast In both cases, a “causal effect” is best understood as some form of *contrast* between the consequences of two or more alternative interventions—*e.g.* taking aspirin as against taking nothing, or taking a placebo. Although such a contrast can be purely qualitative (“Is there a difference or not?”), a more quantitative focus (“How big is the difference?”) is more usual, and more incisive. In many cases the causal variable will

¹A closely related distinction made by philosophers is between “type” and “token” causation.

also be quantitative, leading to questions such as “How much does the effect of aspirin vary with dose?”

Causal variable The above two points lead us to conceive of a *decision variable*, whose values are the various possible interventions under consideration—this is not, however, a “random variable”, since it is up to some agent, rather than Nature, to choose its value.

Relativity We do not delve deeply into such questions as “What is the *true cause*?” of some outcome, but take a pragmatic approach, where the intervention and outcome variables of interest are specified in the light of the context of the problem under consideration. We can change these, as appropriate, to address new questions.

Differences:

Past or future? Typically a EoC query relates to the *future* effect of a past, present, or future action, whereas a CoE query asks about the relationship between *past* actions and outcomes. (However, it is sometimes of interest to consider these last from a historical EoC perspective, looking forward from the time the action was taken, rather than backward from the time the outcome was observed.)

Open or closed? For a EoC problem, where the decision has not yet been implemented, the various choices for the decision variable are all still “open”. In a CoE problem, one value has already been chosen, and the others are “closed”.

What information? More important from a statistical viewpoint than either of the above distinctions is the difference in the the information available for analysis. Specifically, in a CoE situation we typically know—and should therefore take into account—the outcome (of the actually applied action). In an EoC situation the outcome is unknown.

2.3 Hypotheticals and counterfactuals

When we address a EoC query, we are typically asking a *hypothetical* question: “What *would happen* to my headache *if I were to take* aspirin?”. At the very same time we can address alternative hypotheticals: “What *would happen* to my headache *if I were not to take* aspirin?”. In an open situation, where the decision has not yet been made, both questions are meaningful and of real interest, and the “true answer” to either could be observed—if we were to take the relevant action. Ahead of acting, we can try and assess the answer to each (in, say, probabilistic terms), and base our decision on a comparison of these assessments.

The situation is different for a CoE query, where the aspirin has already been taken and the outcome observed. Taking a contrastive view of causality now requires us to address the question: “What *would have happened* to my headache *if I had not taken* aspirin?”. Since the premiss of this question contradicts the known fact that I *did* take aspirin, it is termed a *counterfactual* query. Of logical necessity, the answer to such a question can never be determined. Moreover, if, say, I had taken aspirin and my headache went away remarkably quickly, I would not know whether to ascribe that to the characteristics of the aspirin or of the headache (or some combination of the two). In the former case, the counterfactual duration of my headache (under no aspirin) would be long—in the latter, short.

2.4 Probability and Decision Theory

Let X denote the binary variable denoting whether or not I take aspirin, and Y the log-time it takes for my headache to go away (the actual time will be $Z \equiv e^Y$). We also denote the choice $X = 1$ by t (“treatment”) and $X = 0$ by c (“control”). Although X is a decision variable and so does not have a probability distribution, it is still meaningful to talk of the conditional distribution of Y given $X = x$. Let P_1 [resp. P_0] denote the distribution of Y given $X = 1$ [resp. $X = 0$]. For

the moment we assume these distributions to be known. Where we need to be definite, we shall (purely for simplicity) take these distributions to have the following normal conditional probability density function:

$$p(y | x) = (2\pi)^{-\frac{1}{2}} \exp -\frac{1}{2}(y - \mu_x)^2, \quad (2.1)$$

with a mean μ_0 or μ_1 according as $x = 0$ or 1 , and variance 1 in either case.

The distribution P_1 [P_0] can be interpreted as expressing *hypothetical* uncertainty about Y , if I were to decide on action t [c]. It can incorporate various sources and types of uncertainty, including stochastic effects of external influences arising and acting between the points of treatment application and eventual response.

The distributions P_1 and P_0 are all that is needed to address EoC-type queries: I can simply compare the two different hypothetical distributions for Y , decide which one I would prefer, and take the associated decision.

More formally, we can apply *statistical decision analysis* (Raiffa 1968) to structure and solve this decision problem. Suppose that I quantify the loss I will suffer if my headache lasts y minutes by means of a real-valued loss function, $L(y)$. The decision tree for this problem is as in Figure 2.1. At node ν_1 , $Y \sim P_1$, and the (negative) value of being at ν_1 is measured by the expected loss

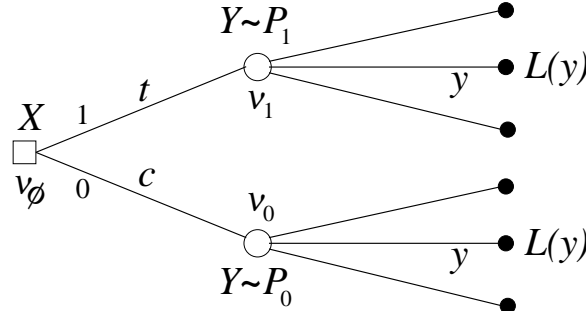


Figure 2.1: Decision tree

$E_{P_1}\{L(Y)\}$. Similarly, ν_0 has value $E_{P_0}\{L(Y)\}$. The principles of statistical decision analysis now require that, at the decision node ν_0 , I choose that treatment, t or c , leading to the smaller expected loss. Whatever loss function is used, this solution will only involve the two hypothetical distributions, P_1 and P_0 .

2.5 Potential responses

However, we can not address a CoE query using only the two distribution P_1 and P_0 —or, equivalently, a conditional probability specification such as (2.1). For suppose I have already taken aspirin ($X = 1$) and observed how long (on the log-scale) it took my headache to go away—say $Y = y$. The relevant counterfactual comparison would relate to my uncertainty about Y under the opposite choice $X = 0$. But I can not condition, at the same time, both on the actual knowledge $X = 1$ and on the counterfactual supposition $X = 0$. More important, I should take account of the actual observation $Y = y$ in assessing what might have happened in different circumstances—but there is nothing in the conditional probability model (2.1) that can support such an analysis.

Thus if we want to address CoE-type queries, it appears we shall need a formalism different from that provided by standard probability theory and decision analysis—one that will allow us to manipulate counterfactuals. The current “industry standard” for this is provided by Rubin’s *potential response* (PR) approach (Rubin 1974; Rubin 1978).

The special feature of the PR approach is that it represents a single response (“effect”) variable Y by two or more random variables—one for each of the possible values of the “cause variable”

X under consideration. Thus, instead of the single response variable Y , “the log-duration of my headache”, the PR approach introduces two *potential response* variables, Y_0 and Y_1 , with Y_x interpreted as “the log-duration of my headache if I take x aspirins”. Both versions Y_0 and Y_1 of Y are conceived as existing simultaneously, ahead of any choice of the value for X . Let \mathbf{Y} denote the pair (Y_0, Y_1) . Then the observed response Y is determined by the values of X and \mathbf{Y} , as $Y = Y_X$.

In this approach, it is assumed that “potential events”, such as “ $Y_0 > Y_1$ ”, have determinate (though of course unknown) truth-values, obeying ordinary 2-valued logic—likewise, “potential variables”, such as Y_1/Y_0 , have determinate numerical values. There is then no formal impediment to assigning a probability distribution to represent joint uncertainty over all potential variables and events.

I like to picture the two potential responses Y_0 and Y_1 as engraved on either side of a “diptych”, *i.e.* a conjoined pair of “tablets of stone” inhabiting Platonic heaven. Each side of the diptych is covered by its own “divine curtain”. When, in this world, X takes value x , just one of these curtains—that associated with the actual value x of X —is lifted, so uncovering, and thereby rendering “actual” and measurable in this world, the associated potential response Y_x . However, because we can only apply one treatment on any occasion, we are never allowed to lift more than one curtain of the diptych.

In contrast to the two univariate distributions P_1, P_0 for Y , we now need to specify a *bivariate* distribution, describing uncertainty about the *pair* of values (Y_0, Y_1) . For consistency with previous assumptions, the marginal distribution of Y_x should be P_x . Note however that those assumptions impose no constraint on the *dependence* between Y_0 and Y_1 . In particular, to be consistent with (2.1) we could assign a bivariate normal distribution to (Y_0, Y_1) , with the given margins, and a correlation coefficient ρ that is simply not determined by our previous assumptions.

The PR framework supplies a formal language in which it is possible to formulate counterfactual speculations, and so address CoE-type queries. Thus if I took an aspirin ($X = 1$) half an hour ago, and my headache has just gone, I can consider the counterfactual probability that, had I not taken aspirin, the headache would still be there—this can be interpreted as the probability that taking the aspirin *caused* my headache to depart when it did. We have already seen that this can not be expressed in terms of the ingredients of the simple conditional distribution model (2.1). It can however be expressed in PR terms, as $\Pr(Z_0 > 30 \mid X = 1, Z_1 = 30)$. The known state of affairs $X = 1$ (aspirin taken), and its observed outcome $Z_1 \equiv e^{Y_1} = 30$, are accounted for by regular conditioning; while interest in the counterfactual state of affairs $X = 0$ (no aspirin taken) is addressed by considering its associated (unobserved) potential response $Z_0 (\equiv e^{Y_0})$. In the PR framework there is no impediment to making both these moves at once.

2.6 What can we know?

This additional expressiveness of the PR machinery may appear liberating. However it is bought at a cost. No matter how we arrange our experiment, it is logically impossible for both values, Y_0 and Y_1 , to be simultaneously observed on the same individual. Consequently, while we might be able to learn, from data, the marginal distribution of either one of these potential responses, there is absolutely no way in which any data could support learning about the *dependence* between them. Any assumptions we might make about this aspect of their bivariate distribution will be just as valid, empirically, as any others; and no data could ever support one set of such assumptions over another.

For example (see §3.2 below) the value of the “probability of causation” $\Pr(Z_0 > 30 \mid X = 1, Z_1 = 30)$ introduced above will vary, depending on the value of the coefficient of correlation ρ between Y_0 and Y_1 . You might assume $\rho = 0$; I might assume $\rho = 1$. So long as we both assign the same marginal distributions to Y_0 and Y_1 , no data that could ever be collected² could say that one of us is right and the other wrong, or even lend more support to one or other of our choices. Consequently we will never be able to choose between the different values that you and I calculate for the probability of causation. So how are we to decide what is its “true value”?

²At any rate, if we do not measure any additional variables.

Chapter 3

Formal Frameworks

In §§ 2.4 and 2.5 we introduced two different formalisms for expressing causality. Here we review these and several variations. In all cases we use the same example of a binary treatment indicator X (aspirin or no aspirin) and a univariate response variable Y (log-duration of headache). For the moment we consider X as a decision variable, and so do not (yet) address the difficulties of making inferences from observational studies.

3.1 Stochastic model

Our simple stochastic model (2.1) just specified the conditional distribution of Y , given either value of X :

$$Y \mid X = x \sim \mathcal{N}(\mu_x, 1). \quad (3.1)$$

This formulation allows us to address EoC-type questions, which only involve comparisons between such distributions. For the simple case of (3.1), where the variance is the same for both treatments, we would typically prefer the treatment with the lower mean. Because it supplies all the ingredients necessary to address prospective decision problems, such as that described by Figure 2.1, we shall also term this the *decision-theoretic* (DT) approach to causality.

However we can not begin to address CoE-type questions using this approach.

3.2 Potential response model

The PR formulation introduces the pair of potential responses, $\mathbf{Y} := (Y_0, Y_1)'$, with a bivariate normal distribution:

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (3.2)$$

Here $\boldsymbol{\mu} := (\mu_0, \mu_1)'$, and $\boldsymbol{\Sigma}$ (2×2) has diagonal entries 1 and off-diagonal entry ρ . It is also implicit that the pair \mathbf{Y} of potential responses is independent of the applied treatment X .

The *observed response* Y is determined by \mathbf{Y} and the treatment indicator X , as $Y = Y_X$. It is easily seen that the conditional distribution of Y given X is given by (3.1)—no matter what the value of ρ may be. So it appears that everything we can do with the stochastic DT model we can also do with the potential response model.

However, we can do more. Thus suppose I took the aspirin ($X = 1$) and the response was $Y = y$. I can pose the counterfactual question: “What would Y have been if I had not taken the aspirin”. In terms of the above ingredients, this can be interpreted as asking about the conditional distribution of Y_0 , given the known data $X = 1, Y = y$ —or, equivalently, given $X = 1, Y_1 = y$. By the assumed independence of \mathbf{Y} from X , this is just the distribution of Y_0 given $Y_1 = y$, which from simple bivariate normal theory is normal with mean and variance:

$$E(Y_0 \mid Y_1 = y) = \lambda \quad := \quad \mu_0 + \rho(y - \mu_1) \quad (3.3)$$

$$\text{var}(Y_0 \mid Y_1 = y) = \delta^2 \quad := \quad (1 - \rho^2). \quad (3.4)$$

By how much did taking the aspirin improve my headache (on the log scale)? This is measured by $y - Y_0$, with relevant distribution $\mathcal{N}(y - \lambda, \delta^2)$. Similarly the actual time improvement is given by $\exp(y) - \exp(Y_0)$, having a shifted log-normal distribution with mean $\exp(y) - \exp(\lambda + \frac{1}{2}\delta^2)$. All such counterfactual and CoE causal questions can be addressed within this framework. But the cost of this increased expressiveness is that the answers to our questions depend on the parameter ρ , which was entirely absent from the DT model (3.1), and can not be identified from data.

3.3 Structural model

Consider again the conditional normal model (2.1). An alternative way in which this would typically be expressed is as follows:

$$Y = \mu_X + E \tag{3.5}$$

where

$$E \sim \mathcal{N}(0, 1) \tag{3.6}$$

and it is implicit that the “error” E is independent of X .

Now certainly the above assumptions are sufficient to imply the distributional property of (2.1). But are they equivalent to it? To assume this is to ignore the additional *algebraic* structure of (3.5), whereby Y is represented as a deterministic mathematical function of the two other variables X and E . Unlike the distributional formulation of (2.1), in (3.5) all the uncertainty is compressed into the single variable E , *via* (3.6). If we take (3.5) and its ingredients seriously, we can get more out of it.

It is implicit in (3.5) that the values of E and X are assigned separately, that of Y then being determined by the equation. Given that E takes value e , then if we set X to x , Y will take value $\mu_x + e$. Thus we can define potential response variables $Y_0 := \mu_0 + E$, $Y_1 := \mu_1 + E$. Being two well-defined functions of the same variable E , they do indeed have simultaneous existence—indeed, they are intimately related, since it is known ahead of time that the difference $Y_1 - Y_0$ will take the non-random value $\mu_1 - \mu_0$. We can thus consider the formulation of the problem in terms of (3.5) and (3.6) as an alternative description of the potential response model of (3.2) for the special case $\rho = 1$.

A system of equations such as (3.5), which may contain hundreds of relationships representing response (“endogenous”) variables as functions of other (both endogenous and “exogenous”) variables as well as of external “error” variables, together with associated explicit or implicit assumptions about the joint distribution of the error terms, constitutes a *Structural Equation Model* (SEM). Such models are particularly popular in econometrics as representations of causal structures—though it is not always entirely clear what assertions about the effects of interventions they are to be understood as making.

3.3.1 Extended structural model

An extension of the structural model (3.5) is given by:

$$Y = \mu_X + E_X \tag{3.7}$$

where we now have two error variables, E_0 and E_1 , with some bivariate distribution—again assumed independent of the value of X . In particular, when $X = 0$ we have $Y = \mu_0 + E_0$, with E_0 still having its initially assigned distribution; and similarly $Y = \mu_1 + E_1$ when $X = 1$. If the marginal distribution of each E_x is standard normal, with density as in (3.6), then (no matter what the correlation ρ between E_0 and E_1 may be) the same distribution model (2.1) for Y given X will be obtained. But again we can go further and define potential responses $Y_x := \mu_x + E_x$, so allowing counterfactual analysis. Conversely, if we start from the potential response model (3.2) we can create an extended structural model on introducing $E_x := Y_x - \mu_x$. The two models are thus essentially identical.

3.4 Functional model

Mathematically, all the models introduced in § 3.2 and § 3.3 have the following common *functional* form:

$$Y = f(X, U), \quad (3.8)$$

where X is a decision variable representing the cause of interest; Y is the effect of interest; U is a further extraneous random variable independent of X ; and f is a deterministic function of its arguments.

In model (3.7), we can take $U = (E_0, E_1)$ and $f(x, (e_0, e_1)) = \mu_x + e_x$; the structural model of (3.5) is the degenerate case of this having $U = E$ and $f(x, e) = \mu_x + e$.

In the case of a potential response model, we can formally take U to be the pair (Y_0, Y_1) , and the function f to be given by:

$$f(x, (y_0, y_1)) = y_x. \quad (3.9)$$

Conversely, any functional model of the general form (3.8) is equivalent to a PR model, if we define $Y_0 = f(0, U)$, $Y_1 = f(1, U)$.¹

We thus see that (mathematically if not necessarily in terms of their interpretation) PR models, extended structural models and general functional models need not be distinguished; while a structural model is a special case of such a model. In particular, all these models support inference about CoE as well as about EoC.

We note that any functional model determines a DT model: under model (3.8) the distribution of Y given $X = x$ is simply the marginal distribution of $f(x, U)$. Conversely, given any DT model $Y | X = x \sim P_x$, we can construct a functional model corresponding it in this way: one simple way is as a potential response model (3.9), in which the marginal distribution of Y_x is P_x . However, in contrast to the essentially unique cross-correspondence between the other models considered above, the functional representation of a DT model is far from unique—as can be seen, for example, from the arbitrariness of the parameter ρ in the PR representation (3.2) of the stochastic model (3.1).

3.5 Fitness for purpose

As discussed in Chapter 2, we can distinguish between causal questions about “effects of causes” (EoC) and those about “causes of effects” (CoE).

DT models, as described in § 3.1 above, can address EoC questions, but can not even express, let alone address, CoE questions.

The other types of model considered above, which are all essentially equivalent, can express both EoC and CoE questions. For CoE, some such more detailed formal framework is essential—although, as shown in § 3.2, its application can be problematic.

It is perhaps of some sociological interest that, notwithstanding the fact that DT models *can* address EoC questions, there is a very widespread perception that the PR framework (or equivalent), with its formalization of counterfactual reasoning, is also essential for handling these problems:

“How is it possible to draw a distinction between causal relations and non-causal associations? In order to meet this concern a further element must be added to the definition—a counterfactual” (Parascandola and Weed 2001).

“Probabilistic causal inference (of which Dawid is an advocate) in observational studies would inevitably require counterfactuals” (Höfler 2005).

¹Any variation in U which does not lead to variation in the pair (Y_0, Y_1) is entirely irrelevant to the relationship between X and Y , so can be ignored.

And indeed, at least 95% of all current research into and discussion of EoC-type causal inference is formulated within the PR framework, simply taking for granted that this is the appropriate formalism to use.

I consider that this is a mistake, and that the simpler DT conception of causality is in fact better suited than PR to address all meaningful problems of EoC. This course will therefore focus on describing and developing the relevant mathematical machinery for use with DT, and on showing how DT can be used to formulate, address and solve a variety of fundamental causal concerns.

Chapter 4

Some Assumptions

4.1 Exchangeability

4.1.1 Simple case

In the simplest variety of statistical inference, we have a “homogeneous population” of individuals. We obtain data by sampling randomly from this population, and use these data to make inferences about the probability structure of the population. Faced with a new individual, who we can also regard as randomly selected from the same population, we consider the probabilistic structure we have learned as applying to him. We can thus use it to make predictions about his as yet unobserved variables, conditional on whatever we have observed on him.

One way of making the above generalities more precise is through the concept of *exchangeability*. This has its origins in the personal/subjectivist world view of Bruno de Finetti (de Finetti 1937; de Finetti 1975), but is much more widely applicable (Dawid 1982; Dawid 1985).

Consider a (typically multivariate) generic variable X , which we can in principle measure for any subject i in the population \mathcal{I} , the specific instance of X for individual i being X_i . A subjectivist statistician could (in principle at least) specify her joint distribution Π for all the $(X_i : i \in \mathcal{I})$. What properties should this joint distribution have?

A classical statistician would naturally model the (X_i) as independent and identically distributed. But this is inappropriate for the subjectivist, who can not leave any aspects of Π numerically unspecified—independence across individuals would entirely preclude the possibility of learning from data on some individuals about the properties of others. It would however often be reasonable to impose the weaker condition that the joint distribution should be unchanged if we considered the individuals in a different order. Thus $(X_5, X_{17}, X_1, \dots)$ should have the same joint distribution as (X_1, X_2, X_3, \dots) , *etc.*—this implies that each X_i has the same marginal distribution, each pair (X_i, X_j) the same bivariate distribution, *etc.* This is the property of exchangeability. Essentially, it just says that the order in which we consider the individuals in the population is entirely irrelevant—thus reflecting, in a generally acceptable way, the intuitions of “homogeneity” and “random sampling” discussed above.

4.1.2 de Finetti’s theorem

The celebrated theorem of de Finetti (1937) constitutes the vital link between the subjectivist and frequentist approaches to Probability and Statistics.

de Finetti showed that (so long at any rate as we can conceive of an unbounded number of exchangeable individuals) exchangeability is mathematically equivalent to the existence of a random distribution P , conditional on which the (X_i) behave as independent and identically distributed draws from P . Here P being “random” means simply that the subjectivist does not know it in advance—that lack of knowledge being expressed by means of probabilities.

On observing data, she will learn more and more about this unknown P , which in turn will enable her to make more useful predictions about new individuals. Given a large enough sample, P will be essentially determined: it is the “objective” distribution of X in the population. Then she would simply predict X_j for a new individual j as following distribution P . With less data, she could take account of the remaining uncertainty about P by taking a further expectation with respect to its posterior distribution.

To remove the personal element, consider now a bevy of subjectivists, having various different joint distributions Π for (X_1, X_2, \dots) , who nevertheless all agree on exchangeability. Then, though they may have different prior opinions about the “objective” distribution P , they will all agree that, given P , the (X_i) are all independent with distribution P . Hence the frequentist’s “independent and identically distributed” model can be considered as the agreed component of all exchangeable distributions: it arises as a necessary consequence of the very weak symmetry assumption of exchangeability. Moreover, given a large enough sample all exchangeable subjectivists will learn the same “true” P (asymptotically equivalent to the empirical distribution in the sample), and use it to describe uncertainty about a new individual.

4.1.3 Conditional exchangeability

Sometimes exchangeability might be an unreasonable assumption, but it can still be assumed that, for some variable Z , all individuals having the same value of Z are exchangeable. This is *conditional exchangeability*. In parallel with the unconditional case, we can show that conditional exchangeability (say for X , given Z) is equivalent to assuming the existence of a (generally unknown) specification $p(x | z)$ of the conditional density¹ function of X given Z ; given which, conditional on $Z_i = z_i$ ($i \in \mathcal{I}$), the (X_i) are independent, X_i having density $p(\cdot | z_i)$.

4.1.4 Causal inference

In a causal setting we can again consider, in an intuitive sense, an exchangeable population of individuals. However we might choose to apply different treatments to different individuals. We shall suppose that this is done in a way that takes no account of the other variables we could measure.

The set \mathbf{U} of “pre-treatment” variables might now be treated as completely exchangeable, and so modelled as independent and identically distributed across the population. All other variables \mathbf{V} (including in particular the response variable Y) might be treated as conditionally exchangeable, *given* treatment T (and \mathbf{U}). Then we can model the situation as follows, in terms of (unknown) densities $p(\mathbf{u})$ and $p(\mathbf{v} | \mathbf{u}, t)$ for (generic) \mathbf{U} , and for \mathbf{V} conditional on receiving treatment t and on \mathbf{U} :

Conditional on treatment assignment,

$$p(\mathbf{u}_i, \mathbf{v}_i | t_i) = p(\mathbf{u}_i) p(\mathbf{v}_i | \mathbf{u}_i, t_i), \quad (4.1)$$

independently across individuals i .

Given data, we can identify $p(\mathbf{u})$ from all individuals, and $p(\mathbf{v} | \mathbf{u}, t)$ from all those receiving treatment t . Having learned these distributions, we can apply them to a new subject, randomly drawn from the same population, to whom we are considering applying one or more of the available treatments.

4.1.5 Extensions

We often wish to extend our causal inferences more boldly than to members of the same population already sampled: perhaps to new populations, or to medical patients who did not meet the specific criteria for inclusion in the study, or even across different species. Or (what will become a major consideration below) our data might have been gathered in a non-random way, *e.g.* because

¹In complete generality densities need not exist, but the general principle still holds.

treatment was not assigned at random, but we nevertheless wish to make causal inferences from our data. Then we can not rely on the above analysis. But we might be able to justify applying these ideas to small components of the overall structure: regarding, say, $p(y | x, t)$ as the same for all individuals involved, be they in our data-base, or intended targets of our predictions. We could then learn these modular components from the data, and combine this partial knowledge with other sources of information to construct our overall predictions.

The property that a conditional distribution $p(y | x)$ is the same across the various different probabilistic “regimes” which we are interested in will be termed *stability* of the distribution of Y given X . When suitable stabilities hold, causal inference can flow much more smoothly; but stability should not be assumed without good cause, and needs to be justified so far as is possible. When stability properties can not be agreed upon, purported causal inferences will remain subject to dispute.

4.1.6 PR approach

The preceding discussion focused on the DT approach to causal inference. Within the PR framework, we would naturally take the various bivariate vectors $\mathbf{Y}_i \equiv (Y_{i0}, Y_{i1})$, comprising the pair of potential responses on individual i , as exchangeable across individuals. In the presence of further individual pre-treatment characteristics \mathbf{U} , we would extend this exchangeability to the $(\mathbf{U}_i, \mathbf{Y}_i)$.² From de Finetti’s theorem we can thus model the $(\mathbf{U}_i, \mathbf{Y}_i)$ as independent and identically distributed from some (unknown) joint density, say $q(\mathbf{u}, \mathbf{y})$.

A link with the treatment in § 4.1.4 (with $\mathbf{V} \equiv \{Y\}$) can now be made:

$$p(\mathbf{u}) \equiv q(\mathbf{u}) \tag{4.2}$$

$$p(y | \mathbf{u}, t) \equiv q(y_t | \mathbf{u}). \tag{4.3}$$

An apparent advantage of this approach is that we can readily account for non-random treatment assignment, by including an observable “treatment variable” T as part of \mathbf{U} , and defining the observed response Y as Y_T . If T is not independent of the potential response Y_t , then (4.3) need not hold, *i.e.* we can allow for confounding in the data. An apparent disadvantage is the difficulty of incorporating extensions as considered in § 4.1.5.

4.2 Treatment-unit additivity

Treatment-unit additivity (TUA) is an assumption often made within the PR approach. It asserts that, although the values of $\mathbf{Y}_i \equiv (Y_{0i}, Y_{1i})$ may differ from one individual i to another, their difference $Y_{1i} - Y_{0i}$ (the “individual causal effect”, ICE, for individual i) is a constant, τ say, across individuals.³ This is already implicit in the structural model of § 3.3; in the more general PR model of § 3.2 it is equivalent to taking the correlation $\rho = 1$. With $\tau = -\log 2$ (say), TUA says that any headache episode would last exactly half as long if treated with aspirin than if left untreated.

Although TUA concerns the dependence between potential responses, it does have some implications for the associated DT model: it requires that the distributions P_0 and P_1 for Y , associated with either one of the two possible treatments, are identical up to a location shift. If this property is false, then TUA can not hold. For example, if we had different variances for different x in (2.1), we could not have TUA. With equal variances, we could identify τ with the difference of means, $\mu_1 - \mu_0$.

If this location-shift property does hold, then TUA might be assumed—though it is by no means implied. It is one thing to say that, if I were to take aspirin, the log-duration of my headache would be *on average* shorter by $\log 2$ than if I did not. It is quite another to say that

²Any other variable in \mathbf{V} that might be affected by treatment would need to be expanded into a collection of potential variables, one for each treatment. We shall not consider this further here.

³Of course this property depend crucially on a particular scale of measurement: it would not be preserved under a non-linear transformation of Y .

my headache will last *exactly* half as long. For DT purposes it is only the average (*i.e.* stochastic) behaviour that matters.

Assuming TUA appears to be a way of resolving the ambiguity exhibited in § 2.6. However, since there is typically no strong reason to assume it, this “resolution” is in fact illusory.

TUA has no real implications for DT analysis, and we shall not consider it further here.⁴

4.3 Stable unit-treatment value assumption

Another basic assumption in the PR approach (Rubin 1980; Rubin 1986) is the *stable unit-treatment value assumption* (SUTVA), which requires that the potential responses to the various treatments that might be applied to an experimental unit be well-defined. In particular, they can not depend on what treatments are applied to other units. If this can not be assumed, then the whole PR description of the problem, including the definitions of experimental unit, treatment, and potential response, must be revised.

From the DT perspective SUTVA (to the extent that it makes any sense at all) appears inappropriate. There is no good reason to require that two different experiments, in each of which treatment t is applied to unit i (but other units may receive divergent treatments) would necessarily deliver identical responses for unit i : for one thing, this would exclude any unpredictable random component of response to treatment. Fortunately, such an assumption is not required for DT. Instead, we could make the generally reasonable assumption that application of treatments does not destroy the homogeneity of the units, beyond the obvious and important difference that some will now have one treatment, some another. Then we will still have exchangeability of the responses for all units (experimental, or future) receiving the same treatment, and can thus use the experimental data to identify the distribution, P_t , of response within treatment group t —which also expresses our uncertainty about the response Y_j of a new unit j , were it to be given treatment t . We then have all the ingredients we need to set up, and solve, the basic treatment choice problem for the new unit j .

4.4 Is there a “fundamental problem of causal inference”?

Within the PR approach, we conceive of the simultaneous existence of both potential responses, Y_0 and Y_1 , and build models for their distribution, jointly with each other and with other variables of interest. But as a logical consequence of their very definition, we can never observe both Y_0 and Y_1 on the same individual: at least one of them will always be missing.⁵ Since the PR conception of causal effect revolves around the comparison of Y_0 and Y_1 , *e.g.* in terms of the ICE, $Y_1 - Y_0$, which would always be unobservable, this missing data issue appears highly problematic: Holland (1986) called it “the fundamental problem of causal inference”. It is intimately related to the fact that it is impossible to learn from data about the dependence between the potential responses.

Fortunately, many analyses conducted within the PR framework make no use of this dependence structure. For example, if we consider the “average causal effect” $ACE := E(Y_1 - Y_0)$, this can be re-expressed as $E(Y_1) - E(Y_0)$, which can thus be identified knowing only the marginal distributions of Y_0 and Y_1 . Likewise, in the problem of § 2.4, where Y is measured on a log-scale, we could identify the ICE on the original scale, $E(Z_1 - Z_0) = E(e^{Y_1}) - E(e^{Y_0})$. However, we can not so identify other comparisons, such as $E(Z_1/Z_0) = E\{\exp(Y_1 - Y_0)\}$, since this *does* depend on the unlearnable dependence between Y_0 and Y_1 .

Within the DT framework there is nothing missing, so no “fundamental problem”. We might be interested in comparing the expectations of Y (or Z) under the two treatment regimes: the difference of these values would be the ACE. But there is simply no DT counterpart of an

⁴A version of TUA expressed entirely in DT terms is introduced and analysed in Dawid (2000), § 8.1.

⁵It is perhaps no accident that the PR approach to causal inference was introduced and developed by Rubin shortly after he had developed important analyses of the general statistical problem of making inferences in the face of missing data (Rubin 1976), and that the PR approach to causal inference conceives of it as a special application of this technology (van der Laan and Robins 2003).

unlearnable quantity such as $E\{\exp(Y_1 - Y_0)\}$. The very “poverty” of the language of DT prevents us from asking, and purporting to answer, silly questions.

When we move to consider CoE questions, we can not avoid the use of a PR-type framework, and (as seen in §3.2) the fundamental problem of causal inference becomes a real problem. It is usually “solved” by simply assuming a particular relationship between Y_0 and Y_1 , for example TUA as in §4.2; but, while that may remove the ambiguity in our causal answers, it does so at the cost of introducing an entirely arbitrary and generally unjustifiable added ingredient.

Part II

**ALGEBRAIC AND
GRAPHICAL
REPRESENTATIONS**

Chapter 5

Conditional Independence

Properties of *independence* and (especially) *conditional independence* (CI) will prove fundamental to understanding statistical causality. Intuitively, when we say that a random quantity X is independent of another Y (written $X \perp\!\!\!\perp Y$), we mean that the distribution of X given $Y = y$ does not depend on y . When we say that X is independent of Y given Z (written $X \perp\!\!\!\perp Y \mid Z$) we mean that the distribution of X given $(Y, Z) = (y, z)$ depends only on the value z of Z . This motivates our formal definition (where we denote by \mathcal{X} the set of values for X , *etc.*, and P denotes the joint distribution of all variables under consideration):

Definition 5.1 [CONDITIONAL INDEPENDENCE] We say X is *independent of* Y , and write $X \perp\!\!\!\perp Y$, if, for any measurable set $A \subseteq \mathcal{X}$, $P(X \in A \mid Y) = P(X \in A)$ [P -almost surely]. We say X is *conditionally independent of* Y given Z , and write $X \perp\!\!\!\perp Y \mid Z$, if, for any such A , $P(X \in A \mid Y, Z) = P(X \mid Z)$ [P -almost surely]. \square

Independence can be regarded as a special case of conditional independence, with the conditioning variable Z being trivial (*e.g.* constant).

When we need to specify explicitly the underlying joint distribution P we will write *e.g.* $X \perp\!\!\!\perp Y \mid Z [P]$.

5.1 Properties and axioms

Suppose for simplicity that all variables are discrete.¹ Let $p(x, y \mid z)$ denote $P(X = x, Y = y \mid Z = z)$, and let $a(x, z)$, for example, denote an unspecified function of (x, z) ; *etc.* Then $X \perp\!\!\!\perp Y \mid Z$ if and only if any of the following equivalent conditions holds:

- C1a : $p(x \mid y, z) \equiv p(x \mid z)$ if $p(y, z) > 0$
- C1b : $p(x \mid y, z)$ has the form $a(x, z)$ if $p(y, z) > 0$
- C2a : $p(x, y \mid z) \equiv p(x \mid z)p(y \mid z)$ if $p(z) > 0$
- C2b : $p(x, y \mid z)$ has the form $a(x, z)b(y, z)$ if $p(z) > 0$
- C3a : $p(x, y, z) \equiv p(x \mid z)p(y \mid z)p(z)$
- C3b : $p(x, y, z) \equiv p(x, z)p(y, z)/p(z)$ if $p(z) > 0$
- C3c : $p(x, y, z)$ has the form $a(x, z)b(y, z)$.

Among the further general properties of probabilistic conditional independence are the following (Dawid 1979a), which may be verified straightforwardly. We here write $W \preceq Y$ to mean that W is a function of Y .

¹These properties readily extend to continuous quantities, though some care is needed. Thus in full generality C1b requires that, for any measurable set A , there exists a measurable function $\alpha : \mathcal{Z} \rightarrow [0, 1]$ such that $P(X \in A \mid Y, Z) = \alpha(Z)$ [P -almost surely].

P1	“Symmetry”	:	$X \perp\!\!\!\perp Y \mid Z$	\Rightarrow	$Y \perp\!\!\!\perp X \mid Z$
P2		:	$X \perp\!\!\!\perp Y \mid X$		
P3	“Decomposition”	:	$X \perp\!\!\!\perp Y \mid Z, W \preceq Y$	\Rightarrow	$X \perp\!\!\!\perp W \mid Z$
P4	“Weak union”	:	$X \perp\!\!\!\perp Y \mid Z, W \preceq Y$	\Rightarrow	$X \perp\!\!\!\perp Y \mid (W, Z)$
P5	“Contraction”	:	$X \perp\!\!\!\perp Y \mid Z$	}	$\Rightarrow X \perp\!\!\!\perp (Y, W) \mid Z.$
			and		
			$X \perp\!\!\!\perp W \mid (Y, Z)$		

(The descriptive terms are those given by Pearl (1988), Chapter 3).

It is possible to derive many further properties of CI by regarding P1 to P5 as axioms for a logical system, rather than calling on more specific properties of probability distributions. A simple example is the following, which expresses the “nearest neighbour” property of a Markov Chain:

Theorem 5.1 *Suppose*

(i). $X_3 \perp\!\!\!\perp X_1 \mid X_2$

(ii). $X_4 \perp\!\!\!\perp (X_1, X_2) \mid X_3$

(iii). $X_5 \perp\!\!\!\perp (X_1, X_2, X_3) \mid X_4$

Then $X_3 \perp\!\!\!\perp (X_1, X_5) \mid (X_2, X_4).$

Proof. Applying P4 and P1 in turn to (ii), we obtain

$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3), \tag{5.1}$$

while from (i) and P1 we have

$$X_1 \perp\!\!\!\perp X_3 \mid X_2. \tag{5.2}$$

On applying P5 to (5.2) and (5.1), we now deduce

$$X_1 \perp\!\!\!\perp (X_3, X_4) \mid X_2 \tag{5.3}$$

whence, by P4 and P1,

$$X_3 \perp\!\!\!\perp X_1 \mid (X_2, X_4). \tag{5.4}$$

Also, by (iii) and P4 we have

$$X_5 \perp\!\!\!\perp (X_1, X_3) \mid (X_2, X_4) \tag{5.5}$$

and so, by P4 and P1,

$$X_3 \perp\!\!\!\perp X_5 \mid (X_1, X_2, X_4). \tag{5.6}$$

The result now follows on applying P5 to (5.4) and (5.6).

□

Some other results obtainable by manipulating CI properties using only P1–P5 are given in Dawid (1979a). These do not involve any properties of probability other than those expressible purely in terms of the CI relation.

5.2 Further axioms?

Another putative property of CI is:

$$\text{P6 "Intersection": } X \perp\!\!\!\perp Y \mid (Z, W) \text{ and } X \perp\!\!\!\perp Z \mid (Y, W) \Rightarrow X \perp\!\!\!\perp (Y, Z) \mid W.$$

The intuitive argument for P6, for the simple case that W is absent, is as follows. Suppose $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$. Since the distribution of X given (Y, Z) (i) does not depend on Z (when Y is given), and (ii) does not depend on Y (when Z is given), it is tempting to conclude that it does not depend on either Y or Z , and thus $X \perp\!\!\!\perp (Y, Z)$. However, this argument does not work (Dawid 1979b), since there may be information common to Y and Z , which could be relevant to X . For example, let Y and Z be functionally unrelated, but suppose the joint distribution gives probability 1 to the event $Y = Z$. Then, for any X , $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$; but we cannot deduce that $X \perp\!\!\!\perp Y$. But P6 does hold if we impose some additional conditions (Dawid 1980): in particular, if the sample space is discrete and each elementary outcome has positive probability.

There are other general properties of probabilistic CI, not derivable from P1–P5: for example (Studený 1989; Studený 1992), if $X \perp\!\!\!\perp Y \mid (Z, W)$, $Z \perp\!\!\!\perp W \mid X$, $Z \perp\!\!\!\perp W \mid Y$ and $X \perp\!\!\!\perp Y$, then $Z \perp\!\!\!\perp W \mid (X, Y)$, $X \perp\!\!\!\perp Y \mid Z$, $X \perp\!\!\!\perp Y \mid W$ and $Z \perp\!\!\!\perp W$. Indeed there are infinitely many mathematically independent such properties. However, for present purposes we will never need to use any general properties of CI other than P1–P5.

5.3 Extension to non-stochastic variables

In order for Definition 5.1 to make sense we must be able to talk about distributions for X , which thus has to be a random variable; but (subject to appropriate interpretation of the “almost sure” qualification) Y and Z need not be. For example, let (X, Y) have a joint distribution P_θ , depending on the value θ of a non-stochastic parameter-variable Θ . We can then interpret $X \perp\!\!\!\perp Y \mid \Theta$ as expressing the probabilistic independence of X and Y under each P_θ . More interestingly, $X \perp\!\!\!\perp \Theta \mid T$ asserts that the conditional distribution of X given $T = t$ under P_θ is the same for all values θ . Similarly, $T \perp\!\!\!\perp \Theta$ denotes that the marginal distribution of T is the same under any P_θ . When T is a function of X these properties are, respectively, just the definitions of *sufficiency* and of *ancillarity* of the statistic T in the family of distributions for X indexed by Θ .

Another important type of non-stochastic variable is a *decision variable*, whose value is determined by a decision-maker, rather than by Nature.

We must exercise a little care when applying the notation and theory of §5.1 to non-stochastic variables, to ensure that these always appear, explicitly or implicitly, as conditioning variables. For example, in the application to sufficiency, where X and Z are stochastic but $Y \equiv \Theta$ is non-stochastic, the property $X \perp\!\!\!\perp \Theta \mid Z$ can not now be re-expressed in the form of C2 or C3 (a or b); C1b should be qualified as holding whenever $p(z \mid \theta) > 0$; and C1a now serves as a *definition* of $p(x \mid z)$ (now only implicitly conditioned on Θ). Again, when T is sufficient ($X \perp\!\!\!\perp \Theta \mid T$) or ancillary ($T \perp\!\!\!\perp \Theta$), we can not use P1 to deduce $\Theta \perp\!\!\!\perp X \mid T$ or $\Theta \perp\!\!\!\perp T$, since, for example, the latter would assert that the conditional distribution of Θ given $T = t$ does not depend on t —a meaningless statement when Θ is non-stochastic.²

Nevertheless, suitably interpreted, properties P1–P5 *do* still hold (Dawid 1980). In fact any deduction made using them will be valid, so long as, in both premisses and conclusions, no non-stochastic variables appear in the left-most term in a conditional independence statement (we *are* allowed to violate this condition in intermediate steps of an argument). So we can apply P1–P5 freely, even in the presence of non-stochastic variables, so long only as we do not attempt to derive any obviously meaningless assertion.

5.4 Conditional independence as a language for causality

Our conception of a *causal relationship* will be very workaday, but nonetheless fit for most practical (rather than metaphysical) purposes for which this concept is required.

²But we *can* make these deductions if we take a Bayesian position and so treat Θ as stochastic. See Dawid (1979a) for more on this.

Firstly, we shall allow fully *stochastic* dependence relationships: the dependence of some (generally multivariate) variable Y on another such variable X is regarded as embodied in a conditional distribution $p(y | x)$.

However, in order to justify regarding such a dependence as *causal* more is needed. To this end we consider, not just one, but a variety of joint distributions for the variables of interest: we term these *regimes*. The various regimes describe different, but related, real-world situations in which we are interested. For example, we might consider different historical times or geographical settings, or patients in different hospitals. In this context, a very important type of regime is that which results from some external *intervention*, perhaps to force some quantity to take on a particular value.

Although we would not normally expect the same joint distributions of all the variables under the different regimes, certain modular components of those distributions might reasonably be expected to be the same. When the conditional distribution $p(y | x)$ is such an invariant ingredient across regimes, we regard the dependence of Y on X as ‘causal’. This property is of particular interest when one or more of the regimes involves an intervention fixing the value of X , but other cases also arise. In any event, this understanding of causality is an entirely relativistic one, highly specific to the particular set of regimes under consideration.

We can introduce a non-stochastic *regime indicator*, F say, which acts like a parameter to index the various different regimes in play. Then the invariance property of the conditional distribution $p(y | x)$ is equivalent to the property that $p(y | x, F = f)$ is the same for all regimes f ; that is to say, to the conditional independence property $Y \perp\!\!\!\perp F | X$. In this way, causal concepts can be re-expressed, and consequently manipulated, using the language and theory of conditional independence. This equivalence will form the basis of most of our analysis of causality below.

Chapter 6

Directed Acyclic Graphs

In many problems, conditional independence properties can be helpfully represented and manipulated using a variety of graph-theoretic tools. Here we consider *directed acyclic graph* representations of CI.

A *directed graph* $\mathcal{D} = (V, E)$ is defined by means of a set V of *vertices* (or *nodes*) (which in our applications will label a set \mathcal{V} of random variables under consideration) and a set $E \subseteq V \times V$ of (*directed*) *edges*. When $v, w \in V$ and $(v, w) \in E$ we draw an arrow from v to w , thus: $v \rightarrow w$ (or $w \leftarrow v$). In this case we say that v is a *parent* of w , and that w is a *child* of v . The set of parents of v is denoted by $\text{pa}(v)$, and the set of children of v by $\text{ch}(v)$. Also if $A \subseteq V$ we define $\text{pa}(A)$ as $\bigcup_{v \in A} \text{pa}(v)$, etc.

A *directed path* from v to w is a sequence $v = v_1 \rightarrow v_2 \dots \rightarrow v_n = w$. When such a path exists we write $v \mapsto w$, and call v an *ancestor* of w , w a *descendant* of v (we also regard any node as both an ancestor and a descendant of itself). The set of ancestors of v is denoted by $\text{an}(v)$, the set of descendants of v by $\text{de}(v)$, these definitions again extending to sets of nodes on taking unions. The set of non-descendants of v is denoted by $\text{nd}(v)$.

The directed graph is *acyclic* if, for every $v \in V$, there is no directed path from v to v . Then we have a *directed acyclic graph* (DAG).

6.1 DAG representation of a distribution

Let $\mathcal{V} = \{X_v : v \in V\}$ be a labelled set of variables, with joint distribution P , and suppose the label-set V has been ordered somehow. We can then identify V with the sequence $(1, 2, \dots, N)$. Correspondingly we have the ordered sequence of variables $\mathbf{X} = (X_1, \dots, X_N)$. For $1 \leq v \leq N$ let $\text{pre}(v) = (1, \dots, v-1)$ (and $\text{pre}(0) = \emptyset$), and let $X_{\text{pre}(v)}$ denote the subsequence (X_1, \dots, X_{v-1}) .

We now construct a DAG \mathcal{D} , with vertex-set V , as follows.

Introduce each node v in order. For $v = 1, \dots, N$, consider the distribution of X_v given $X_{\text{pre}(v)}$. Let $\text{pa}(v)$ be defined as a subset of $\text{pre}(v)$ such that this conditional distribution in fact depends only on $X_{\text{pa}(v)}$. (There is always such a subset, if only the full set $\text{pre}(v)$; but we would usually want $\text{pa}(v)$ to be chosen as small as possible). We then draw an arrow from each $w \in \text{pa}(v)$ to v . In the resulting DAG $\text{pa}(v)$ is indeed the set of parents of v .

We note that, for any $v \in V$, the defining property of the set $\text{pa}(v)$ can be expressed in terms of conditional independence:

$$X_v \perp\!\!\!\perp X_{\text{pre}(v)} \mid X_{\text{pa}(v)}. \quad (6.1)$$

Example 6.1 Suppose that P is a distribution for (X, Y, Z) such that $X \perp\!\!\!\perp Z \mid Y$. If we take the variables in the order (X, Y, Z) , the corresponding graph is $X \rightarrow Y \rightarrow Z$; for the order (Z, Y, X) , it is $X \leftarrow Y \leftarrow Z$; for the orders (Y, X, Z) or (Y, Z, X) it is $X \leftarrow Y \rightarrow Z$; while for any other

order it is a complete graph, with arrows between each pair of vertices.¹ \square

6.2 Factorization of the density

We can decompose any joint density $p(\mathbf{x})$ of $\mathbf{X} = (X_1, \dots, X_n)$ in terms of successive conditional densities, as follows:

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v \mid x_{\text{pre}(v)}).$$

When (6.1) holds, this simplifies to:

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v \mid x_{\text{pa}(v)}). \quad (6.2)$$

For a given DAG $\mathcal{D} = (V, E)$, and distribution P for $\mathcal{V} = \{X_v : v \in V\}$, we say that P *factorizes according to \mathcal{D}* if (6.2) holds for its density. For any well-ordering of V (i.e. such that $v \rightarrow w \Rightarrow v < w$), this is equivalent to requiring (6.1). In particular, this property is entirely determined by the graph, and not the way in which its nodes are numbered.

6.2.1 Ancestral sets

A set $A \subseteq V$ of nodes of a DAG is termed *ancestral* if, whenever $v \in A$ and $u \rightarrow v$, then $u \in A$. Then A must contain all the ancestors of any of its members.

Lemma 6.1 *Suppose the density $p(\mathbf{x})$ factorizes according to a DAG \mathcal{D} , as in (6.2). Let A be an ancestral set. Then the marginal joint density of x_A is*

$$p(x_A) = \prod_{v \in A} p(x_v \mid x_{\text{pa}(v)}). \quad (6.3)$$

In particular, $p(x_A)$ factorizes according to \mathcal{D}_A , the *induced subgraph* of \mathcal{D} over A , whose vertex set is A , and whose edges are just those edges in E both of whose endpoints are in A .

Proof. If $A = V$ there is nothing to prove. Otherwise there must be at least one childless node w in $V \setminus A$. Then x_w will occur in the right-hand side of (6.2) only in the term $p(x_w \mid x_{\text{pa}(w)})$, and integrating out over x_w simply removes this term. Hence the result holds for $A = V \setminus \{w\}$. Continuing in this way, we eventually obtain the desired result for any ancestral A . \square

6.3 Conditional independence properties implied by a DAG

Suppose that the distribution P factorizes according to the DAG \mathcal{D} . For any node v , we can always well-order the set V such that $\text{pre}(v) = \text{nd}(v)$. It follows that we must have

$$X_v \perp\!\!\!\perp X_{\text{nd}(v)} \mid X_{\text{pa}(v)}. \quad (6.4)$$

This property is expressed by saying that P is *directed Markov* with respect to \mathcal{D} . It is fully equivalent to the factorization property (6.2). But still other conditional properties are implied by DAG-factorization.

¹In special cases there may be further (conditional) independences implied by P , leading to deletion of some of the arrows in these graphs; but this will not happen in general.

6.4 Moralization

To explore these, we define the *moral graph* of a DAG \mathcal{D} to be the undirected graph \mathcal{D}^m obtained by first “marrying unmarried parents”, *i.e.* adding undirected edges between any pair of parents of a vertex which are not already joined by an edge, and finally making all edges undirected.

For subsets A, B, S of V , we say S *separates A from B in \mathcal{D}^m* if, in that graph, every path from a node in A to a node in B intersects S .

Theorem 6.2 *Suppose that P factorizes according to \mathcal{D} , and that S separates A from B in \mathcal{D}^m . Then $X_A \perp\!\!\!\perp X_B \mid X_S [P]$.*

Proof. Let $A^* := \{v \in V : S \text{ separates } v \text{ from } B\} \setminus S$, $B^* := V \setminus (A^* \cup S)$. Then A^*, B^*, S form a partition of V , $A \subseteq A^* \cup S$, $B \subseteq B^* \cup S$, and S separates A^* from B^* .

Suppose a term of the form $p(x_v \mid x_{\text{pa}(v)})$ involves both x_a and x_b , for some $a \in A^*, b \in B^*$. There are three possible cases to consider: $v = a$ and $b \rightarrow a$; $v = b$ and $a \rightarrow b$; or $a \rightarrow v$ and $b \rightarrow v$. In any of these cases, a and b are neighbours in the moral graph \mathcal{D}^m . But this is impossible, since S separates A^* from B^* in \mathcal{D}^m . Consequently, any term $p(x_v \mid x_{\text{pa}(v)})$ is a function either of $x_{A^* \cup S}$ or of $x_{B^* \cup S}$.

On collecting terms together, it now follows that (6.2) can be expressed in the form

$$p(\mathbf{x}) = f(x_{A^*}, x_S) g(x_{B^*}, x_S). \quad (6.5)$$

Thus, from (C3c) of §5.1, $X_{A^*} \perp\!\!\!\perp X_{B^*} \mid X_S [P]$. Finally, on applying P1–P5 (and using $X_A \perp\!\!\!\perp (X_{A^*}, X_S)$, *etc.*), $X_A \perp\!\!\!\perp X_B \mid X_S [P]$ follows. \square

This result can be extended, as follows.

Suppose P factorizes according to \mathcal{D} , and let $A, B, S \subseteq V$. Consider $V' := \text{an}(A \cup B \cup S)$, the smallest ancestral set containing A , B and S , and denote the moralized graph of $\mathcal{D}_{V'}$ by $\text{man}(A \cup B \cup S)$. On using Lemma 6.1, we can apply Theorem 6.2 to the marginal distribution of $X_{V'}$ to deduce the following.

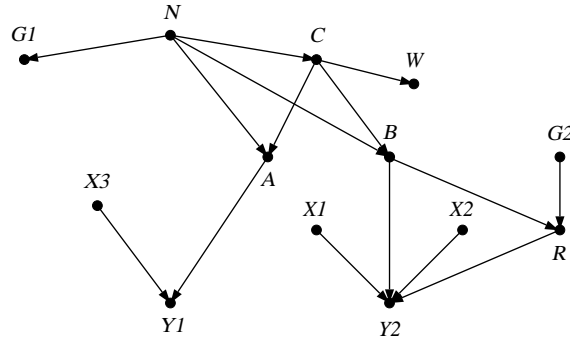
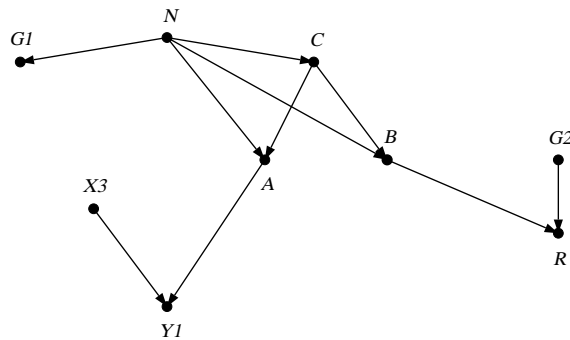
Definition 6.1 We write $A \perp_{\mathcal{D}} B \mid S$ to denote the property that S separates A from B in $\text{man}(A \cup B \cup S)$. \square

Theorem 6.3 (Moralization criterion) *Suppose that P factorizes according to \mathcal{D} , and $A \perp_{\mathcal{D}} B \mid S$. Then $X_A \perp\!\!\!\perp X_B \mid X_S [P]$.*

This property is also expressed by saying that P satisfies the (*global*) *directed Markov* property with respect to \mathcal{D} .

It can be shown that the purely graph-theoretic separation property $\perp_{\mathcal{D}}$ obeys essentially the same formal properties P1–P5 (see §5.1) as probabilistic CI. An alternative, more abstract, proof of Theorem 6.3 can be based on this and the easily verified fact (exactly parallel to (6.4)) that, for any DAG \mathcal{D} , $X_v \perp_{\mathcal{D}} X_{\text{nd}(v)} \mid X_{\text{pa}(v)}$ (Lauritzen *et al.* 1990).

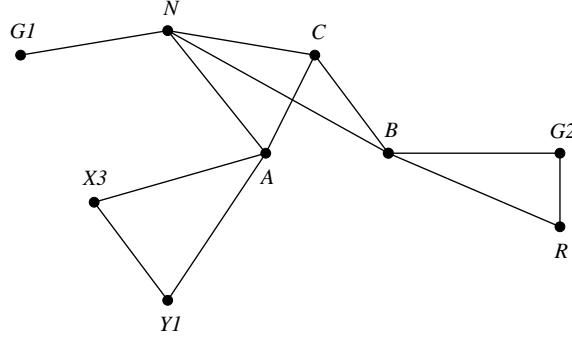
It can further be shown (Geiger and Pearl 1990) that the conditional independencies one can deduce by applying Theorem 6.3 exhaust all those of the form $X_A \perp\!\!\!\perp X_B \mid X_S$ that hold for every distribution P that factorizes according to \mathcal{D} , which are in turn all those that can be formally deduced by repeatedly applying P1–P5 to the input conditional independence properties of (6.1). Thus Theorem 6.3 can be used as a fully powerful “theorem-proving machine”, allowing one easily to read off all the consequences of one’s initial conditional independence assumptions (so long as these can be represented as a list of the form (6.1)). For example, the unrevealing algebraic proof of Theorem 5.1 can be replaced by simple inspection of the DAG $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$ that embodies the assumptions made.

Figure 6.1: Directed graph \mathcal{D} for criminal evidenceFigure 6.2: Ancestral subgraph \mathcal{D}'

Example 6.2 The graph \mathcal{D} of Figure 6.1 describes the relationships between the evidence and other variables figuring in a criminal trial (Dawid and Evett 1997). The graph is constructed so that the local directed Markov property captures the assumed dependence structure. For example, the distribution of $Y1$ (measured properties of a tuft of fibres found at the scene), given all other variables, is supposed fully determined by the values of $X3$ (properties of the suspect’s jumper) and of A (an indicator of whether or not the fibres came from the suspect’s jumper). Thus, even before attempting numerical specification of the joint distribution, we know that it will be directed Markov over \mathcal{D} . We can therefore read off certain conditional independence properties that the distribution must possess. For instance, we can show that $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$ as follows. First we construct the ancestral subgraph $\mathcal{D}' = \mathcal{D}_{\text{an}(B, R, G1, Y1, A, N)}$, containing the vertices of interest and all their ancestors (Figure 6.2). We then moralize it, by “marrying unmarried parents” and dropping the arrowheads, to obtain $\mathcal{G}' = (\mathcal{D}')^m$ (Figure 6.3).

We now note that, in \mathcal{G}' , it is impossible to trace a path from either of B or R to either $G1$ or $Y1$ without it intersecting the set $\{A, N\}$, *i.e.* $(B, R) \perp_{\mathcal{D}} (G1, Y1) \mid (A, N)$. Since the joint distribution is directed Markov over \mathcal{D} , we can thus deduce the probabilistic conditional independency $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$. (Properties such as these have been used to simplify expressions for the likelihood ratio in favour of guilt in the light of the evidence.) \square

Caution: It should be remarked that, although every DAG describes some collection of conditional independence properties, and can be used to manipulate these, by no means every such collection can be represented by a DAG. Thus while DAG models can be very helpful when they can be used, they are not always available. In full generality, we may need to use algebraic manipulations, applying the CI axioms P1–P5 to derive the implicit consequences of any assumed collection of conditional independencies.

Figure 6.3: Moralized ancestral subgraph \mathcal{G}'

6.5 d -separation

We shall use the moralization criterion of Theorem 6.3 to derive CI properties implied by a DAG representation of a distribution. However there is an alternative formulation that is commonly seen, based on the property of d -separation (Pearl 1986; Verma and Pearl 1990). This we now describe.

The *skeleton* \mathcal{D}^\sim of a DAG \mathcal{D} is the undirected (but unmoralized) graph obtained from \mathcal{D} by ignoring the directions of the arrows. A *trail* in \mathcal{D} is a sequence of distinct vertices forming a path in \mathcal{D}^\sim . A node v of a trail π is HH if it is an internal node of π , and the arrows of π meet head-to-head at v .

Let $S \subseteq V$, A trail π is said to be *blocked* (by S) if it contains a node c such that *either*

- $c \in S$ and c is not HH; *or*
- c and all its descendants are not in S , and c is HH.

A trail that is not blocked by S is said to be *active*. Two subsets A and B are said to be d -separated by S if all trails from A to B are blocked by S .

Theorem 6.4 *Suppose the distribution P is directed Markov with respect to \mathcal{D} . Then if A , B , and S are pairwise disjoint sets of vertices, $A \perp\!\!\!\perp B \mid S [P]$ whenever S d -separates A from B in \mathcal{D} .*

Theorem 6.4 can be proved directly (Verma and Pearl 1990). Alternatively, it follows from Theorem 6.3 and the following purely graph-theoretic result.

Theorem 6.5 (Lauritzen *et al.* (1990); Richardson and Spirtes (2002), Theorem 3.18) *Let A , B , and S be pairwise disjoint subsets of a directed acyclic graph \mathcal{D} . Then S d -separates A from B if and only if $A \perp_{\mathcal{D}} B \mid S$.*

Proof.² Let $W := \text{an}(A \cup B \cup S)$, let \mathcal{D}' denote \mathcal{D}_W , and let \mathcal{G}' be the moralization of \mathcal{D}' .

Suppose S does not d -separate A from B . Then there is an active trail π from A to B such as, for example, the one indicated in Figure 6.4.

All internal nodes c in this trail must lie within W . For if c is HH, then $c \in S$ or c has descendants in S . If c is not HH, we can start to follow a subpath directed away from c until we either meet an incoming arrow, necessarily at a HH node, in which case c has a descendant in S ; or until we get to a or b , in which case c has a descendant in A or B . Hence $\pi \subseteq \text{an}(A \cup B \cup S)$.

If an internal node v of π is in S , then v must be HH. Since neither of its neighbours, say α , β , in π can also have this property, they must both lie outside S . There must be a link $\alpha \rightarrow v \leftarrow \beta$ in the moral graph \mathcal{G}' , as illustrated in Figure 6.5. On replacing every such configuration $\alpha \rightarrow v \leftarrow \beta$ by the undirected edge $\alpha - \beta$, and dropping remaining arrows, we create a path from A to B in \mathcal{G}' , circumventing S . Consequently we do not have $A \perp_{\mathcal{D}} B \mid S$.

²This proof fills some gaps in that given by Lauritzen *et al.* (1990) and repeated in Cowell *et al.* (1999). I am grateful to Thomas Richardson for pointing these out.

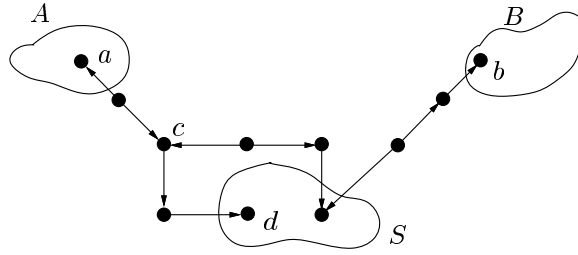


Figure 6.4: Example of an active trail from A to B . (The path from c to d is not part of the trail, but indicates that c must have descendants in S .)

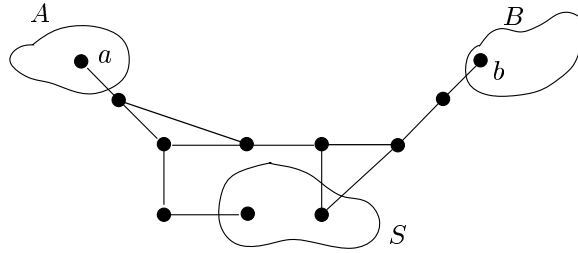


Figure 6.5: The moral graph corresponding to the active trail in \mathcal{D} .

Conversely, suppose we do not have $A \perp_{\mathcal{D}} B \mid S$. Let \mathcal{T} denote the set of paths from A to B in \mathcal{G}' that circumvent S , and whose interior nodes all lie outside $A \cup B$; then \mathcal{T} is non-empty. The edges of any such path either correspond to directed edges of the original DAG \mathcal{D} , or are “moral edges” $\alpha-\beta$ created by marriages of “immorality” configurations $\alpha \rightarrow \gamma \leftarrow \beta$ in \mathcal{D} , where $\gamma \in W$.

Let $\pi^{\sim} \in \mathcal{T}$ with the smallest possible number of moral edges. Starting from π^{\sim} we can create a path³ π in \mathcal{D}' by replacing each moral edge $\alpha-\beta$ by its originating immorality $\alpha \rightarrow \gamma \leftarrow \beta$, and restoring the direction of the arrow on any other edge. We will show that π is an active trail.

Lemma 6.6 *The path π is a trail (i.e., all its nodes are distinct).*

Proof. Let a, b be the end-points of π^{\sim} in A and B respectively.

Suppose $\alpha \rightarrow \gamma \leftarrow \beta$ is an originating immorality as above, and without loss of generality suppose that α lies between a and β in the path π^{\sim} . Let \mathcal{A} denote the subpath of π^{\sim} from a to α (inclusive), and similarly \mathcal{B} that from β to b .

We first remark that we can not have $\gamma \in A$. For if so, the path $\gamma-\mathcal{B}$, where the connecting edge derives from $\gamma \leftarrow \beta$, would be in \mathcal{T} but have fewer moral edges than π^{\sim} —a contradiction. Similarly $\gamma \notin B$.

Now suppose $\gamma \in \pi^{\sim}$. Without loss of generality let $\gamma \in \mathcal{A}$, and let \mathcal{A}_{γ} denote the subpath of π^{\sim} from a to γ . Then the concatenated path $\mathcal{A}_{\gamma}-\mathcal{B}$, where again the connecting edge derives from $\gamma \leftarrow \beta$, would be in \mathcal{T} but with fewer moral edges than π^{\sim} —a contradiction. So $\gamma \notin \pi^{\sim}$.

Finally suppose we have two such immoralities, $\alpha_1 \rightarrow \gamma_1 \leftarrow \beta_1$ and $\alpha_2 \rightarrow \gamma_2 \leftarrow \beta_2$; without loss of generality let α_1 be closer than α_2 to a in the path π^{\sim} . If $\gamma_1 = \gamma_2$, then $\alpha_1-\beta_2$ is an edge (possibly originally directed, but if not then moral) in \mathcal{G}' , and replacing the original portion of π^{\sim} between α_1 and β_2 by this edge would again produce a path in \mathcal{T} with fewer moral edges than π^{\sim} . Hence γ_1 and γ_2 must be distinct. \square

Every node of the trail π that is not in π^{\sim} must have been introduced as the result of a marriage. In particular, since π^{\sim} avoids S , any node in π that is in S must be HH. To show that π is active, we therefore need to show that, if $c \in \pi$ is HH, then either c or one of its descendants is in S .

Suppose not. Then c must have descendants in A or B —for definiteness, say in A . Let δ be a shortest directed path in \mathcal{D} leading from c into A (see Figure 6.6), and denote by λ the part of the trail π between c and b . Then δ and λ intersect at c , and might also intersect elsewhere. Let w be “lowest” node in δ that is also in λ . Concatenating the part of δ leading from w into A and the part of λ leading from w into B now forms a trail κ in \mathcal{D}' having at least one less HH node than π : Figure 6.6

³i.e., a sequence of nodes such that, for any consecutive nodes a, b either $a \rightarrow b$ or $a \leftarrow b$.

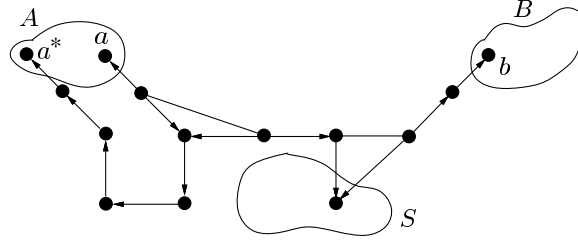


Figure 6.6: The path π^\sim in $\text{man}(A \cup B \cup S)$ makes it possible to construct an active trail π from A to B in \mathcal{D} .

show this for the case $w = c$. From κ we can now construct a path in \mathcal{T} having fewer moral edges than π^\sim . But since this would contradict the definition of π^\sim , we conclude that π is active, and hence that S does not d -separate A from B . \square

6.6 Markov equivalence

We have seen that, associated with any DAG \mathcal{D} , there is a collection of conditional independence properties, as described by Theorem 6.3 or equivalently Theorem 6.4, that hold for all distributions P that are directed Markov with respect to \mathcal{D} . However, distinct DAGs can represent identical collections of conditional independencies—in which case they are termed *Markov equivalent*.

Recall that the *skeleton* of a DAG \mathcal{D} is the undirected graph \mathcal{D}^\sim obtained by ignoring the directions of the arrows on the edges of \mathcal{D} . An *immorality*, or *v-structure*, in \mathcal{D} is a configuration of the form $a \rightarrow c \leftarrow b$, where a and b are parents of a common child c but are “unmarried”, *i.e.* neither $a \rightarrow b$ nor $b \rightarrow a$.

Theorem 6.7 (Frydenberg (1990); Verma and Pearl (1991)) *Two DAGs \mathcal{D}_0 and \mathcal{D}_1 on the same vertex set V are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

Example 6.3 There are just three possible DAGs on two nodes:

- (i). $A \rightarrow B$
- (ii). $A \leftarrow B$
- (iii). $A \quad B$.

Since DAGs (i) and (ii) have the same skeleton, and neither has any immoralities, they are Markov equivalent: indeed, they embody no conditional independence properties whatsoever. Any joint density $p(a, b)$ can be factorized either according to (i), as $p(a)p(b | a)$ or, equivalently, according to (ii), as $p(b)p(a | b)$.

However, DAG (iii), which has a different skeleton, embodies the non-trivial conditional independence restriction $A \perp\!\!\!\perp B$. \square

Example 6.4 Consider the following DAGs on three nodes:

- (i). $A \rightarrow B \rightarrow C$
- (ii). $A \leftarrow B \leftarrow C$
- (iii). $A \leftarrow B \rightarrow C$
- (iv). $A \rightarrow B \leftarrow C$.

These all have the same skeleton. However, whereas DAGs (i), (ii) and (iii) have no immoralities, (iv) has one immorality. Consequently, (i), (ii) and (iii) are Markov equivalent to each other, but (iv) is not Markov equivalent to these. Indeed, (i), (ii) and (iii) all express the conditional independence property $A \perp\!\!\!\perp C \mid B$, whereas (iv) expresses the marginal independence property $A \perp\!\!\!\perp C$. \square

6.7 Influence diagrams

We have seen how certain collections of stochastic conditional independence properties can be represented by DAGs, where each vertex of the DAG is associated with a random variable.

In §§ 5.3 and 5.4 we extended stochastic conditional independence to allow some of the variables to be parameter, regime or decision variables. Correspondingly we can extend DAGs to do the same. What results is an *influence diagram* (ID) (Howard and Matheson 1984; Shachter 1986; Oliver and Smith 1990; Lauritzen and Nilsson 1999; Nilsson and Lauritzen 2000), a DAG in which some of the nodes are designated as *stochastic* or *random*, and marked by a circle or oval, and others are non-stochastic (typically, though not invariably, interpreted as *decision nodes*), being marked by a square or rectangle.

Example 6.5 In Figure 6.7, nodes B , D and E are random nodes. Associated with each random node is its conditional distribution given its parent nodes, just as in a fully probabilistic DAG.

Nodes A and C are decision nodes. The value of a decision variable is determined by the outside intervention of a decision maker, DM, rather than being left to arise naturally. The parents of a decision node represent the information that is supposed available to DM, at the point at which that decision has to be made.

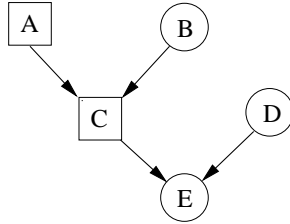


Figure 6.7: Influence diagram

A full description of a particular pattern of behaviour by DM would require specification, for each decision node, of a function of the values of its parent nodes, describing exactly which of the available decisions is to be selected for any particular set of information. We can extend this further to allow randomized decision-making, described by specifying conditional probability distributions over the available decisions at a node, given its parents. A given specification, π say, of these functions or distributions at decision nodes constitutes a *decision strategy*. Combining this with the given conditional distributions at random nodes, we then recover exactly the structure of a probabilistic DAG—and this applies, albeit in degenerate form, also in the non-randomized case.

We thus suppose that we have specified externally the densities $p(b)$, $p(d)$, and $p(e \mid c, d)$; while DM can choose a value for (or distribution over) the decision at A ; and, for each pair of values of A and B , a value for (or distribution over) the decision at C .

The extra strategy information required to turn an influence diagram into a DAG is not part of the basic specification: conditional distributions are given for random nodes, but the functions or distributions involved at the decision nodes are arbitrary and at the choice of the decision maker. The specified inputs determine what we may term the *partial* distribution of random nodes, given decision nodes. In our example, this is

$$p(b, d, e : a, c) = p(b)p(d)p(e \mid c, d).$$

Similarly, any decision strategy π specifies a partial distribution, this time for the decision nodes given the random nodes: in our example this is

$$\pi(a, c : b, d, e) = \pi(a)\pi(c | a, b),$$

where the terms are to be interpreted as probabilities or densities. Once a strategy π has been specified we readily obtain the full joint distribution for all the variables:

$$p_{\pi}(a, b, c, d, e) = p(b, d, e : a, c)\pi(a, c : b, d, e). \quad (6.6)$$

It is important to realize that the partial distributions are not, in general, the same as the corresponding conditional distributions calculated from the full joint distribution. \square

Chapter 7

Causal Interpretations Of DAGs

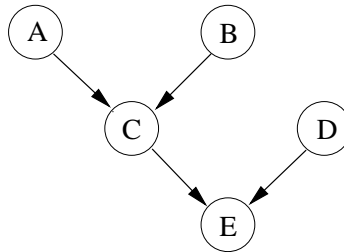


Figure 7.1: A probabilistic DAG

The DAG of Figure 7.1, like any DAG, can be considered as a representation of properties of conditional independence between variables associated with its vertices. Specifically these are: $A \perp\!\!\!\perp B$; $D \perp\!\!\!\perp (A, B, C)$; and $E \perp\!\!\!\perp (A, B) \mid (C, D)$; together with all other properties, such as $E \perp\!\!\!\perp B \mid (A, C)$, deducible from the above using P1–P5, or, equivalently, readable off the DAG using the moralization criterion of Theorem 6.3 or the d -separation criterion of Theorem 6.4.

From one point of view, the *only* use for a DAG such as \mathcal{D} is to express such conditional independence properties—using the pure *stochastic DAG semantics* already described. In particular, purely graph-theoretic concepts such as arrow, parent, child should not be interpreted as referring to anything in the outside world: they are merely internal cogs in a technical machinery for expressing conditional independence (a relation that, it should be noticed, does not incorporate any concept of directionality, whereas a DAG does). In particular, using these semantics we would have no reason to distinguish between Markov equivalent DAGs, such as those described in Examples 6.3 and 6.4.

However, going beyond the above purely stochastic semantics, it is common to interpret an arrow $a \rightarrow b$ in a DAG as representing some kind of “direct causal dependence” of X_b on X_a . This vague statement can be (and often is) left vague, or else clarified by means of an extended “causal DAG semantics”.

7.1 Intervention DAGs

One such semantics (Pearl 2000, Definition 1.3.1) treats a DAG such as that of Figure 7.1 as representing, not merely a joint distribution under naturally observed conditions, but also various different joint distributions that might arise as a result of external interventions in the system. Thus consider a “point” (or “atomic”) intervention, that forces one of the variables in the system, say C , to take on a pre-assigned value, say c_0 . Since this is an entirely new state of affairs from the non-interventional situation described by Figure 7.1, there is no reason to expect any relationship

whatsoever between the joint distributions of the variables in the two cases. But under an extended “causal” interpretation of Figure 7.1, such a relationship is assumed. Specifically, it is supposed that the new distribution still factorizes according to Figure 7.1, with the identical specification for the parent-child distributions $p(x_v | x_{\text{pa}(v)})$ *except for* $v = C$, when this is replaced by the one-point distribution on the assigned value c_0 . In particular, since this new conditional distribution does not in fact depend on the parents (A and B) of C in Figure 7.1, node C , together with all its incoming arrows, is now redundant and could be removed. The distributions of variables which are non-descendants of C (namely A , B and D) are entirely unaffected; while that of E given D becomes $p(e | c_0, d)$. Thus the joint distribution of (A, B, D, E) when C is set to c_0 is supposed given by

$$p(a, b, d, e | C \leftarrow c_0) = p(a)p(b)p(d)p(e | c_0, d). \quad (7.1)$$

Here the notation ‘ $C \leftarrow c_0$ ’ indicates ‘conditioning by intervention’, in contradistinction to ‘ $C = c_0$ ’, which we reserve for ordinary conditioning; alternative notations that have been used are $p(a, b, d, e | \check{c}_0)$, $p(a, b, d, e | \hat{c}_0)$, $p(a, b, d, e || C = c_0)$, and $p(a, b, d | \text{do}(C = c_0))$.

Taken to its limits (which is how Pearl and others appear to interpret it), the above procedure could be applied to determine a revised joint distribution for the variables under a point intervention at any one of the nodes, or simultaneous point interventions at any set of nodes. Hence under this interpretation the DAG model of Figure 7.1 is regarded as modelling, in closely related ways, all the various distributions of the variables, in a whole variety of distinct circumstances. We call a DAG equipped with these modified semantics an *intervention DAG*.¹

Caution: It is important to appreciate that whether or not such a description of the effects of interventions is appropriate can never be a purely mathematical question, but must be assessed in the light of what is in fact the case in the empirical world that the intervention DAG is purporting to model. In particular this needs to take account of the real-world interpretations of the mathematical terms. For example, there will usually be a variety of mechanisms by which the value of a variable could be set: setting a patient’s aspirin treatment to ‘none’ by (a) withholding it from him, (b) wiring his jaw shut, or (c) killing him are all very different interventions, with different effects, and we must be very clear as to which mechanism we are considering. An intervention DAG model can be justified only to the extent that it fits the behaviour of the world in the setting to which it is intended to apply. Furthermore, since the main question is whether or not the various interventional situations are indeed related to the non-interventional one in the specific way described by the DAG, no assessment of its appropriateness can be made (at any rate, on any purely empirical basis) if we only have data from the non-interventional distribution.

Note that we can not calculate (7.1) if all we know is the initial joint density of the variables: we also need to know the assumed DAG structure. In particular, distinct DAGs that are Markov equivalent as probabilistic DAGs are no longer equivalent when interpreted as intervention DAGs, since they will yield different intervention formulae.

Example 7.1 The DAGs of Figure 7.2 and Figure 7.3, which (see Example 6.3) are equivalent as probabilistic DAGs, are no longer equivalent when interpreted as intervention DAGs. Thus after an intervention to set A to a_0 , the new distribution of B would be its original *conditional* distribution, $p(b | a_0)$, for Figure 7.2, but its original *marginal* distribution, $p(b)$, for Figure 7.3. Thus A ‘affects’ B in the former model, but not in the latter.

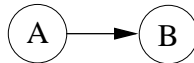


Figure 7.2: A before B

¹Lauritzen (2000) uses this description for what we shall term below an *augmented DAG*. Pearl (2000) calls our intervention DAG a *causal Bayesian network*.

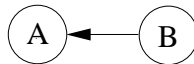


Figure 7.3: B before A

We will never be able to decide which, if either, of these distinct causal models is appropriate merely by examining observational data: to do this we need to perform experiments in which we do in fact intervene to set the value of A , and note what effect that intervention has on B . And it may well turn out that neither of the above choices represents the actual distribution of B in the experiment—in which case the intervention DAG semantics are simply inappropriate to describe the actual causal relationships. \square

7.1.1 Seeing and doing

On making the appropriate replacement for $p(c | a, b)$ in the product formula (6.2), we readily see that the joint interventional density of all the variables, at configuration $\mathbf{x} = (a, b, c, d, e)$, is

$$p(\mathbf{x} | C \leftarrow c_0) = \frac{p(\mathbf{x})}{p(C = c_0 | A = a, B = b)} \quad (7.2)$$

so long as $c = c_0$ (and is otherwise 0).

This should be contrasted with the effect of *conditioning*, in the purely observational regime, on the observation that $C = c_0$. In that case we have, instead,

$$p(\mathbf{x} | C = c_0) = \frac{p(\mathbf{x})}{p(C = c_0)} \quad (7.3)$$

for $c = c_0$ (and otherwise 0). Since C is not independent of (A, B) in the non-interventional joint distribution p , the effect (7.2) of *doing* $C = c_0$ is thus different in general from the effect (7.3) of *seeing* $C = c_0$. The effect of seeing can be calculated knowing only the observational joint density; whereas that of doing requires further knowledge of the structure of an intervention DAG assumed to describe the causal structure.

7.2 Augmented DAGs

Having two different semantics linked to the same graphical representation (either a purely probabilistic DAG, representing conditional independence in a fixed regime, or an intervention DAG, representing relationships between distributions in different regimes) is clumsy, and can lead to confusion. Here we show how an extended graphical representation for the causal setting, using an influence diagram rather than a simple DAG, can make this distinction explicit and clear.

Let X be a random variable taking values in \mathcal{X} . Suppose that it is possible to intervene somehow to force X to take on some chosen value $x \in \mathcal{X}$. We now represent this explicitly by introducing an *intervention variable* F_X , which is a special kind of decision variable. The state space of F_X is constructed by augmenting \mathcal{X} with an additional state \emptyset , also termed “idle”. Intuitively, if $F_X = \emptyset$, X is allowed to arise ‘naturally’; while a value $x \in \mathcal{X}$ for F_X indicates an intervention to set the value of X to x . In particular, the conditional distribution of X , given $F_X = x \in \mathcal{X}$ (and any other information) will be degenerate at the value x ; while the conditional distribution of X given $F_X = \emptyset$, and any other information H , will be just its ‘natural’ distribution, given H .

Formal conditioning on $F_X = \emptyset$ is supposed to have no effect on any other variables in the problem. However, the global effect of an intervention, $F_X = x$ (beyond its immediate effect at setting $X = x$) needs further specification. There is no magic solution to this problem: what is appropriate must depend on the context and meaning of the variables (including the way in which intervention is effected), and on what assumptions appear reasonable in the circumstances.

The semantics of an intervention DAG, as described in § 7.1, provide one way of modelling and describing global response to intervention. An alternative, more explicit and versatile, representation of these assumptions is by means of an ‘augmented DAG’: a special kind of influence diagram, which, for each existing domain node X of the original probabilistic DAG, explicitly adds a new intervention node F_X , as described above, as a decision node parent of X (Spirtes *et al.* 1993; Pearl 2000; Lauritzen 2000).

Formally, let $\text{pa}^0(X)$ be the set of *domain parents* of X , *i.e.* those random variables featuring as parents of X in the interventional DAG. In the augmented DAG (which is now an influence diagram, ID), a decision node representing the intervention variable F_X is explicitly introduced, and becomes another parent of X . The conditional distribution of X , given any configuration \mathbf{y} of $\text{pa}^0(X)$ and the value x of F_X , is taken to be the same as the original distribution for X given $\text{pa}^0(X) = \mathbf{y}$ if $x = \emptyset$; but otherwise puts all its probability mass on the value x for X .

It is easily seen that, as represented by the augmented DAG, the conditional joint distribution of all the domain variables given $F_X = x$ ² is just the joint observational distribution when $x = \emptyset$, and otherwise agrees with the distribution implied by applying the semantics of § 7.1 to the unaugmented DAG \mathcal{D} to model the effect of an intervention setting X to x . This readily extends to conditioning on a set of intervention nodes (all others being idle). Consequently, for any sets of domain variables A, B, C , $\Pr(A = a \mid B = b, F_C = c)$, as calculated from the augmented DAG, is the same as $\Pr(A = a \mid B = b, C \leftarrow c)$ as calculated by applying intervention semantics to the unaugmented DAG \mathcal{D} .

Once we have constructed the augmented DAG, we can apply the full formal machinery of moralization (see § 6.4) or equivalently d -separation (see § 6.5) to it; but because we are now able to include the decision variables together with the domain variables, we can also use DAG-separation properties to express causal, as well as purely probabilistic, assertions, as described in § 5.4.

Example 7.2 The augmented DAGs corresponding to Figures 7.2 and 7.3 (regarded as intervention DAGs) are given by the influence diagrams of Figures 7.4 and 7.5.

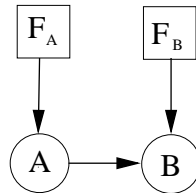


Figure 7.4: Augmented DAG: A causes B

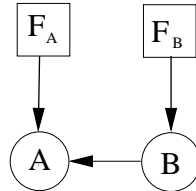


Figure 7.5: Augmented DAG: B causes A

An immediate payoff of the explicit display of the intervention variables in the DAGs is that we can see directly that these graphs do not represent equivalent causal assumptions since, although they have the same skeleton, they have different immoralities. Furthermore, using *e.g.*

²Here and throughout, any unmentioned regime indicator is to be regarded as idle.

the moralization criterion, we see that Figure 7.2 expresses the conditional independence property $B \perp\!\!\!\perp F_A \mid A$ ³: this says that the *conditional* distribution of B given $A = a$ is the same no matter what the value of F_A may be, *i.e.* the same in the interventional regime, when $F_A = a$, as in the observational regime, when $F_A = \emptyset$. However Figure 7.3 expresses graphically the conditional independence property $B \perp\!\!\!\perp F_A$, and thus makes it explicit that it the *marginal* distribution of B is the same, no matter whether (and how) A is subjected to intervention, or not. \square

We thus see how the implicit causal properties of an intervention DAG are rendered explicit, and simply expressible by graphical representations of conditional independence, when we elaborate it into an augmented DAG.

Example 7.3 The augmented DAG corresponding to Figure 7.1 is given by Figure 7.6.

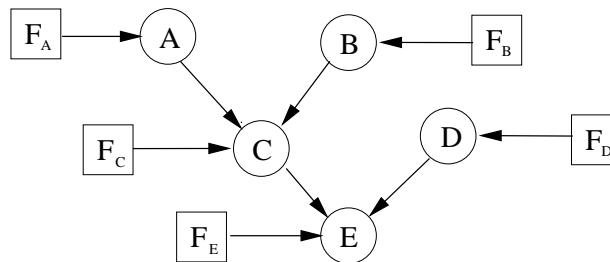


Figure 7.6: Augmented DAG

We can read off the graph that, for example, $C \perp\!\!\!\perp (D, F_A, F_B, F_D, F_E) \mid (A, B, F_C)$: conditional on its parents, C is conditionally independent of its other non-descendant nodes. In particular, if we fix F_C at \emptyset (and drop this conditioning from the notation), the ‘natural’ conditional distribution of C given A and B is not further affected by additional conditioning on the value of D , *nor by whether or not any or all of A, B, D or E arose naturally or by intervention*. Similar properties hold for any other domain node in place of C . In particular the conditional distribution for a node, given its domain parents, when it is allowed to ‘arise naturally’, remains unchanged when its parents are set by intervention. That is, the augmented DAG *explicitly* encodes the assumptions that are only implicit in the intervention interpretation of a probabilistic DAG; and further makes it possible to read off their implications directly. \square

An augmented DAG model expresses exactly the same assumed structure as an intervention DAG model; but the augmented DAG representation is to be preferred, since the assumed structure is explicitly represented in the diagram, its implications can easily be read off the graph, and equivalence of different structures is easy to check. (Note however that the requirement that $p(X \mid F_X = x, \text{pa}^0(X))$ be degenerate at x ($x \neq \emptyset$) is not explicitly displayed in the graph, and still has to be introduced as a separate externally specified constraint).

7.2.1 Extensions

The augmented DAG representation also easily allows for the incorporation of additional flexibility, by varying specific details while retaining the general structure of an influence diagram. For example, we could restrict the possibility of intervention to a subset of the nodes, or to a subset of the state-space of a node; or modify the effect of intervention, for example so as to determine the value of the associated node in some externally decided random way; or allow dependence between the decision nodes, or have a decision node dependent on ancestral nodes in the graph, or make some domain nodes pure decision nodes. All such modifications are easily made explicit and manipulated by means of suitable influence diagrams, such as that of Figure 6.7.

³Since F_B is also a decision variable we should also be conditioning on that: for a non-trivial result we take it as set to its observational value $F_B = \emptyset$.

7.3 Functional DAGs

An alternative graphical representation for probabilistic and causal structures uses *functional models*, based on functional rather than probabilistic dependence relations: these form the basis of Pearl's later approach to causal modelling (misleadingly, he calls them 'probabilistic causal models'). We introduce these in the purely probabilistic setting; causal extensions return in §7.4 below.

Consider the *functional DAG* of Figure 7.7. Here the observable domain variables are A and B ,

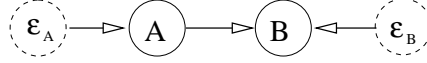


Figure 7.7: Functional DAG, A causes B

while ε_A and ε_B represent independent unmeasured 'exogenous random errors', indicated by dashed circles. The hollow arrows are used to indicate externally specified *deterministic* dependencies, given by functional relations of the form:

$$A = g_A(\varepsilon_A), \quad (7.4)$$

$$B = g_B(A, \varepsilon_B). \quad (7.5)$$

These can also be considered as special degenerate forms for the conditional probability specifications at nodes A and B , so that Figure 7.7 is indeed a DAG representation of the joint probabilistic structure of $(\varepsilon_A, \varepsilon_B, A, B)$.

In order to complete the description of the model, we need to specify the dependence of each node in Figure 7.7 on its parents: that is to say, the marginal probability distributions of ε_A and ε_B , and the functional relationships g_A and g_B in (7.4) and (7.5). Having done this we can obtain the joint distribution of $(\varepsilon_A, \varepsilon_B, A, B)$, and thus the joint distribution for (A, B) by marginalizing over $(\varepsilon_A, \varepsilon_B)$.

It is not hard to see that, by suitable choice of the distributions of ε_A and ε_B and the functions g_A, g_B , we can obtain any desired joint distribution for (A, B) in this way. Thus suppose that A, B are real-valued (extensions to the multivariate case are straightforward), with a joint distribution represented by the DAG model of Figure 7.2 (which is to say, with no restriction whatsoever). We can define the full probability structure by specifying the distribution function G_A of A , and, for each value a of A , the conditional distribution function G_B^a of B given $A = a$. Define ε_A and ε_B to be independently uniform on $[0, 1]$, $g_A(\cdot) \equiv G_A^{-1}(\cdot)$, and $g_B(a, \cdot) \equiv (G_B^a)^{-1}(\cdot)$. It is then easily verified that the implied marginal distribution of A , and the conditional distributions for B given A , are exactly as desired. It is also easy to see that this is just one of many distinct ways in which we could specify a functional DAG that induces the desired distributions for A , and for B given A .

We can similarly represent *any* probabilistic DAG by means of a functional DAG, constructed by adding a new exogenous 'error node' ε_v (again represented by a dashed circle) feeding into each domain node v , different error variables being taken as mutually independent. Thus a functional DAG representing the probabilistic DAG of Figure 7.1 would look like Figure 7.8, with, again, the functional

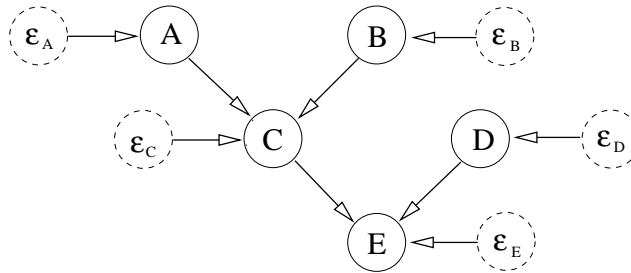


Figure 7.8: Functional DAG

relationships and the distributions of the error variables externally specified. It is readily seen, using the moralization criterion, that the conditional independence properties between the domain variables embodied in Figure 7.8 are identical with those embodied in Figure 7.1. And once again, for each node v , we can reproduce *any* desired specification of $p(v \mid \text{pa}^0(v))$ by suitable choice of the distribution of the corresponding error variable ε_v and the function g_v in $v = g_v(\text{pa}^0(v), \varepsilon_v)$. Consequently we can recreate an arbitrary specification of a probabilistic DAG structure by starting from some functional

model, in which all the randomness is confined to the error variables. We reiterate, however, that the details of such functional representations are far from unique. The following example (essentially recapitulating §3.3 above) makes this clear.

Example 7.4 Consider the following joint distribution associated with the probabilistic DAG of Figure 7.2. A is a binary variable, with

$$\Pr(A = 1) = 1 - \Pr(A = 0) = p. \quad (7.6)$$

Also, conditionally on $A = a$ ($a = 0, 1$), B has the normal distribution

$$B \sim \mathcal{N}(\mu_a, 1) \quad (7.7)$$

for given μ_0, μ_1 .

We now introduce $\boldsymbol{\varepsilon}_B = (\varepsilon_{B,0}, \varepsilon_{B,1})$, having a bivariate normal distribution with each margin $\mathcal{N}(0, 1)$, and correlation ρ . Define g_B by: $g_B(a, \boldsymbol{\varepsilon}_B) = \mu_a + \varepsilon_{B,a}$. We can complete the functional description, introducing an additional variable ε_A and equation $A = g_A(\varepsilon_A)$, in a variety of ways, to recreate the marginal distribution of A : the specific details of this are not relevant to our current purpose. We then obtain a functional representation, as described by Figure 7.7, of our original problem. Note particularly that the desired conditional distribution of B given $A = a$, viz. $\mathcal{N}(\mu_a, 1)$ ($a = 0, 1$), will follow from (7.5), *no matter what may be the value assigned to ρ* . \square

We thus see that there can be distinct functional models which represent the same probabilistic DAG. In the above example, the differences are confined to the form of the distribution for $\boldsymbol{\varepsilon}_B$. In general we could also vary the functional relationships. These arbitrarinesses are over and above the variety of ways by which the same joint distribution may be represented by different probabilistic DAG diagrams, as in the case of Figures 7.2 and 7.3.

7.3.1 Latent variable models

A functional DAG is a special case of a *DAG with unmeasured (latent) variables*—in this case, the ε 's. A more general case arises when we do not insist that the arrows in a diagram such as Figure 7.8 be hollow, *i.e.* the dependence of C on (A, B, ε_C) , *etc.*, is allowed to be probabilistic rather than deterministic. In this case too, the implied joint distribution for the original domain variables (A, B, C, D, E) will have the independence properties expressed in Figure 7.1.⁴ This gives yet more ways in which we can represent that original structure—by introducing, and then ignoring, additional variables. And once again, there is a very wide variety of ways in which this could be done. In certain problems (see Chapter 11) the additional variables introduced might themselves be unmeasured domain variables, with clear external meaning, and in such cases it could indeed be helpful to expand the DAG in this way. However, when the additional variables are pure mathematical fictions, introduced merely so as to reproduce the desired probabilistic structure of the domain variables, there seems absolutely no good reason to include them in the model. Moreover there is a danger that, if we incorporate into our model terms which do not have any clear external referent, we may all too easily lose sight of this fact, and attempt to make inference about them, or impose additional conditions on them, so leading us to conclusions that might appear to be meaningful (because they can be expressed mathematically in terms of the ingredients we have chosen to include in our model), but which in fact have no scientific basis (since those ingredients are themselves largely arbitrary, and this arbitrariness may feed through to our ‘conclusions’).

7.4 Functional intervention models

Just as we were able to extend the semantics of a probabilistic DAG to model certain assumption about the effects of interventions, so we can introduce an intervention semantics into a functional model. In the probabilistic case we needed to know the various conditional probability distributions associated with each node, and describe the effect of interventions on these. Now we need to know the functional relationships and error distributions, and describe how interventions are supposed to affect these. Thus consider the effect of ‘setting’ C to c_0 in Figure 7.8. One way of modelling this is as follows (though we again emphasize that this is by no means the only possibility, and whether or not it provides a good model must remain an empirical question). We suppose that, on setting C to c_0 , the original functional relation $C = g_C(A, B, \varepsilon_C)$ is now deleted, to be replaced by $C = c_0$; while *all other functional relations remain unchanged*.

To complete the specification, we still need to describe the effect of the intervention on the error variables. The assumption we shall make is that there is *no effect* of intervention on the (joint) *distribution* of the errors. We may describe this by saying that the errors are supposed *insensitive* to intervention. Note that this is different from the assumption usually made, that the actual *values*

⁴In general, however, if we marginalize out over certain variables in a DAG model, the joint distribution over the remainder may not itself be describable in terms of any DAG.

of the errors are unchanged by intervention, a situation that may be described as *unresponsiveness*. In full generality neither condition need imply the other, although in a functional intervention model unresponsiveness implies insensitivity (Heckerman and Shachter 1995).

Just as for a probabilistic intervention DAG, the assumptions implicit in a functional intervention DAG can be represented more explicitly and usefully by its augmented form, incorporating an additional decision node F_v for each domain node v . The augmented functional intervention DAG corresponding to Figure 7.7 is as in Figure 7.9.

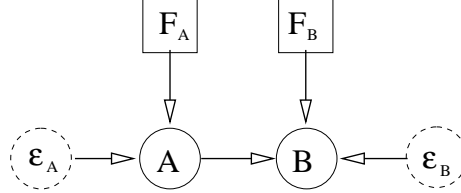


Figure 7.9: Augmented functional DAG, A causes B

Equation (7.5) is now replaced by:

$$B = h_B(A, \varepsilon_B, F_B), \quad (7.8)$$

where $h_B(a, u, \emptyset) \equiv g_B(a, u)$, while, for $x \neq \emptyset$, $h_B(a, u, x) \equiv x$. Equation (7.4) is adjusted similarly. Apart from the fact that we never associate interventions with the error nodes, the augmented functional DAG is related to the functional DAG (which is after all just a special probabilistic DAG) in exactly the way we have described in §7.2. Note that the assumption that the ε 's be insensitive to (*i.e.* independent of) the F 's is encoded explicitly, by the standard DAG semantics, in Figure 7.9.

Again, we can readily represent variations in the nature and scope of interventions. For example, the 'functional influence diagram' of Figure 7.10, together with Equation (7.5), represents a situation

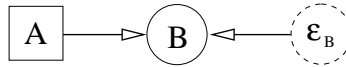


Figure 7.10: Functional influence diagram, A causes B

in which A is a pure decision node, entirely under experimental control, while no direct intervention on B is possible. (Indeed, this more restricted structure was all that was really needed for our description of Example 7.4).

It is easy to see that, if we start with a functional DAG representation of a probabilistic DAG, then the above defined functional intervention DAG is likewise equivalent to the associated probabilistic intervention DAG, in the sense that the probabilistic consequences of any intervention will be the same in both descriptions. Similarly, any probabilistic influence diagram, such as that of Figure 6.7, can be represented by means of a *functional influence diagram*, having, for each random domain node, a new exogenous error node as an additional parent. The distributions and functions can be chosen (though, we again note, non-uniquely) to represent any desired partial distribution for random nodes given decision nodes.

Part III

SPECIAL TOPICS

Chapter 8

Computing Causal Effects

In the presence of unmeasured variables it may or may not be possible to identify causal effects from data on the remainder.

Example 8.1 Consider a problem represented by the augmented DAG of Figure 8.1, where the dashed circle indicates that U is unmeasured (and intervention at such a variable, and in this case at Y , is not envisaged).

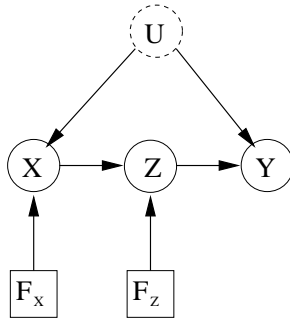


Figure 8.1: A DAG with unmeasured variables

Using *e.g.* moralization one readily checks the following conditional independence properties between the measured variables (each statement being meaningful even though the intervention variables F_X , F_Z are non-stochastic):

$$Y \perp\!\!\!\perp F_Z \mid (X, F_X, Z) \quad (8.1)$$

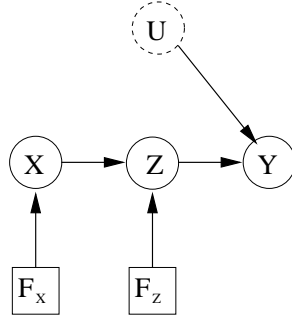
$$Z \perp\!\!\!\perp F_X \mid (X, F_Z) \quad (8.2)$$

$$X \perp\!\!\!\perp F_Z \mid F_X. \quad (8.3)$$

These do not exhaust all the conditional independence properties implied by the assumptions of the augmented DAG. In particular, whenever any variable is set by intervention, its dependence on its domain parents is voided, and we can delete all the associated arrows from the DAG. For example, if X is being set, the problem can now be represented by Figure 8.2, from which we can read off the following further conditional independence:

$$Y \perp\!\!\!\perp (X, F_X, F_Z) \mid (Z, F_X \neq \emptyset). \quad (8.4)$$

Yet another property that holds for any measured variable V in any augmented DAG is: $F_V = v$ ($v \neq \emptyset$) $\Rightarrow V = v$. However this can not be expressed solely in terms of conditional independence or graphs.

Figure 8.2: DAG when X is set ($F_X \neq \emptyset$)

We can attempt to apply the above properties to compute a target causal distribution, such as $p(y | F_X = x)$ ($x \neq \emptyset$), from distributions conditioned on $F_X = F_Z = \emptyset$, which are identifiable under observational conditions. For simplicity, we assume a discrete joint distribution in the following. Any unmentioned intervention variable is assumed set at \emptyset .

Suppose first we want the causal effect of X on Z , *i.e.* $p(z | F_x = x)$ ($x \neq \emptyset$). Since $F_X = x \Rightarrow X = x$, this is equivalent to $p(z | X = x, F_X = x)$. But from (8.2) $Z \perp\!\!\!\perp F_X | X$, so that this is the same as $p(z | X = x)$ —and so identifiable from observational data.

Next consider the causal effect of Z on Y , $p(y | F_Z = z)$. Now we do not necessarily have $Y \perp\!\!\!\perp F_Z | Z$, so this is not immediately identifiable. However (using $F_Z = z \Rightarrow Z = z$) we can express

$$p(y | F_Z = z) = \sum_x p(y | X = x, Z = z, F_Z = z) p(X = x | F_Z = z). \quad (8.5)$$

By (8.1) $Y \perp\!\!\!\perp F_Z | (X, Z)$, whence the first term on the right-hand side is $p(y | X = x, Z = z)$. Similarly, from (8.3) $X \perp\!\!\!\perp F_Z$, so the second term is $p(X = x)$. Since both terms can thus be identified¹ in the observational regime, so can $p(y | F_Z = z)$.

Finally, consider the causal effect of X on Y , $p(y | F_X = x)$. We have:

$$p(y | F_X = x) = \sum_z p(y | X = x, F_X = x, Z = z) p(Z = z | X = x, F_X = x). \quad (8.6)$$

From (8.2) the second term is $p(Z = z | X = x)$, and so observationally identifiable.

As for the first term, by (8.1) this is $p(y | X = x, F_X = x, Z = z, F_Z = z)$. Now applying (8.4) (further conditioned on F_Z), this reduces to $p(y | F_Z = z)$, which we have already shown to be identifiable by formula (8.5). So we can identify (8.6). \square

8.1 General approach

We here make use of the notation of Pearl (2000) in which *e.g.* $p(y | x, \check{z})$ refers to $\Pr(Y = y | X = x, F_Z = z)$ (it being implicit that $z \neq \emptyset$, and all unmentioned intervention variables are idle.)

Let X, Y, Z, W be arbitrary sets of variables in a problem also involving intervention variables. The following rules follow² from the very definition of conditional independence.

Rule 1 (Insertion/deletion of observations) If $Y \perp\!\!\!\perp Z | (X, F_X \neq \emptyset, W)$ then

$$p(y | \check{x}, z, w) = p(y | \check{x}, w). \quad (8.7)$$

¹To be fully rigorous we need to impose further “positivity conditions” to ensure that the desired conditional distributions are well-defined. Thus for the term $p(y | X = x, Z = z)$ to be well-defined we must require $p(z | X = x) > 0$ —at any rate for the relevant x -values, *i.e.* those for which $p(X = x) > 0$.

²Again, we shall ignore throughout the further positivity conditions required to ensure that the relevant conditional probabilities are well-defined.

Rule 2 (Action/observation exchange) If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, Z, W)$, then

$$p(y \mid \tilde{x}, \tilde{z}, w) = p(y \mid \tilde{x}, z, w). \quad (8.8)$$

Rule 3 (Insertion/deletion of actions) If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, W)$, then

$$p(y \mid \tilde{x}, \tilde{z}, w) = p(y \mid \tilde{x}, w). \quad (8.9)$$

Example 8.1 illustrated how successive application of these rules, coupled with the property $F_X = x \Rightarrow X = x$ and the laws of probability, can sometimes allow one to express a "causal" expression in purely observational terms. In particular, the argument there for $p(y \mid F_Z = z)$ can be expressed in the following completely general terms (where simply to be confusing we have interchanged X and Z):

Theorem 8.1 (Back-door formula) *Suppose that*

$$Z \perp\!\!\!\perp F_X \quad (8.10)$$

$$Y \perp\!\!\!\perp F_X \mid (X, Z). \quad (8.11)$$

Then

$$p(y \mid F_X = x) = \sum_z p(y \mid Z = z, X = x) p(Z = z). \quad (8.12)$$

Similarly the argument in Example 8.1 for $p(y \mid F_X = x)$ can be expressed generally as:

Theorem 8.2 (Front-door formula) *Suppose that*

$$Y \perp\!\!\!\perp F_Z \quad (8.13)$$

$$Z \perp\!\!\!\perp F_X \mid X \quad (8.14)$$

$$Y \perp\!\!\!\perp F_Z \mid (X, Z, F_X) \quad (8.15)$$

$$Y \perp\!\!\!\perp (X, F_X) \mid (Z, F_Z, F_X \neq \emptyset) \quad (8.16)$$

Then

$$p(y \mid F_X = x) = \sum_x p(z \mid x) \sum_{x'} p(y \mid x', z) p(x'). \quad (8.17)$$

8.2 DAG models and Pearl's "do calculus"

Suppose now that our problem is represented by an augmented DAG \mathcal{D} . Then (just as we did in Example 8.1) we can use the moralization criterion to check the conditions for applying Rules 1–3.

Let $\mathcal{D}_{\overline{X}}$ denote the "manipulated" augmented DAG obtained from \mathcal{D} on removing all arrows into any node in X from its domain parents. Under intervention at X (*i.e.*, $F_X \neq \emptyset$) the problem is represented by $\mathcal{D}_{\overline{X}}$, which is therefore the relevant graph to use to query any of the above rules. Thus we have:

Rule 1 If $Y \perp_{\mathcal{D}_{\overline{X}}} Z \mid (X, W)$, then $p(y \mid \tilde{x}, z, w) = p(y \mid \tilde{x}, w)$.³

Rule 2 If $Y \perp_{\mathcal{D}_{\overline{X}}} F_Z \mid (X, Z, W)$, then $p(y \mid \tilde{x}, \tilde{z}, w) = p(y \mid \tilde{x}, z, w)$.

Rule 3 If $Y \perp_{\mathcal{D}_{\overline{X}}} F_Z \mid (X, W)$, then $p(y \mid \tilde{x}, \tilde{z}, w) = p(y \mid \tilde{x}, w)$.

³In fact in this setting Rule 1 is redundant, since it is easily seen that $Y \perp_{\mathcal{D}_{\overline{X}}} Z \mid (X, W)$ implies both $Y \perp_{\mathcal{D}_{\overline{X}}} F_Z \mid (X, Z, W)$ and $Y \perp_{\mathcal{D}_{\overline{X}}} F_Z \mid (X, W)$. So when this condition holds we can use Rule 2 to replace $p(y \mid \tilde{x}, z, w)$ by $p(y \mid \tilde{x}, \tilde{z}, w)$, followed by Rule 3 to replace this by $p(y \mid \tilde{x}, w)$.

Pearl (1995a) gives alternative expressions for the above separation properties in $\mathcal{D}_{\overline{X}}$, purely in terms of the underlying intervention DAG $\mathcal{G}_{\overline{X}}$ obtained from $\mathcal{D}_{\overline{X}}$ by removing the intervention nodes and related arrows. For the case that the sets X, Y, Z, W are disjoint, he shows that these conditions are equivalent to:

Rule 1 $Y \perp_{\mathcal{G}_{\overline{X}}} Z \mid (X, W)$.

Rule 2 $Y \perp_{\mathcal{G}_{\overline{XZ}}} Z \mid (X, W)$, where $\mathcal{G}_{\overline{XZ}}$ is obtained from $\mathcal{G}_{\overline{X}}$ on removing all arrows out of nodes in Z .

Rule 3 $Y \perp_{\mathcal{G}_{\overline{X \cup Z(W)}}} Z \mid (X, W)$, where $Z(W)$ is the set of nodes in Z that have no descendants in W . (The proof of equivalence, credited to David Galles, is somewhat intricate).

For any of these cases the relevant graphical separation condition can be checked either by moralization or by d -separation in the relevant unaugmented graph. But the previous approach of querying the augmented graph $\mathcal{D}_{\overline{X}}$ directly is generally more straightforward.

Pearl terms the application of the above rules the “*do*-calculus”. In the context of an augmented DAG model, when only expressions involving measured variables are considered, it can be shown (Shpitser and Pearl 2006a; Shpitser and Pearl 2006b; Huang and Valtorta 2006) that this calculus is *complete*, in the sense that, whenever there exists a reduction of a causal expression to observational terms, it can be constructed by such successive application of the above three rules (together with the condition $F_X = x \Rightarrow X = x$, and the general laws of probability).

8.2.1 Back-door and front-door

Pearl interprets (8.10) and (8.11) in terms of d -separation properties of the unmanipulated intervention graph \mathcal{G} , as follows:

$$Z \subseteq \text{nd}(X) \quad (\text{in } \mathcal{G}) \tag{8.18}$$

$$Z \text{ blocks all back-door trails from } X \text{ to } Y \text{ in } \mathcal{G}. \tag{8.19}$$

Here a *back-door trail* from X to Y means one that leaves X by an arrow directed into X (hence the description of Theorem 8.1).

Similarly, he argues that conditions (8.13)—(8.16) of Theorem 8.2 will hold when the following properties hold in the intervention DAG \mathcal{G} :

- (i). Z intercepts all directed paths from X to Y .
- (ii). There is no back-door trail from X to Z .
- (iii). All back-door trails from Z to Y are blocked by X .

8.3 Nonidentifiability of causal effect

In certain circumstances we can show that a causal effect can *not* be identified from observational data (at any rate, without imposing further assumptions).

8.3.1 Bow-pattern

Definition 8.1 Let \mathcal{D} be an augmented DAG, with measured domain variables \mathcal{M} and unmeasured domain variables \mathcal{U} . Let $X, Y \in \mathcal{M}$. A *bow-pattern* between X and Y is a trail from X to Y in the ancestral graph $\text{an}_{\mathcal{D}}(X, Y)$ of $\{X, Y\}$ in \mathcal{D} , whose arrows at the endpoints are directed into X and into Y , and all of whose intermediate nodes are in \mathcal{U} . An essentially identical definition applies to an intervention graph. \square

Remark 8.1 In the presence of a bow-pattern, there will be a path from F_X to Y in $\text{man}_{\mathcal{D}}(X, Y)$ all of whose intermediate nodes are in \mathcal{U} .

Remark 8.2 A bow-pattern in \mathcal{D} will be retained in any manipulated DAG resulting from intervening at any $Z \subseteq \mathcal{M} \setminus \{X, Y\}$.

We can indicate the presence of a bow pattern by a double-headed arrow, as in Figure 8.3 which derives from Figure 8.1.

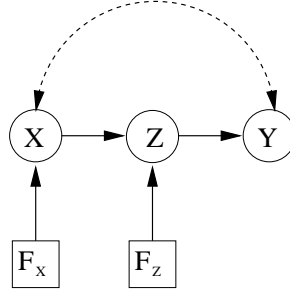


Figure 8.3: A bow pattern

A special case of a bow-pattern occurs when there exists $U \in \mathcal{U}$ that is an ancestor of both X and Y , and directed paths from U to X and Y passing through only unmeasured variables: this is the only case considered by Pearl (2000) (§ 3.5).

8.3.2 Parent-child bow

Suppose that $Y \in \text{ch}(X)$, and there is a bow-pattern between X and Y , as in Figure 8.4.

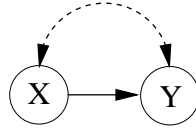


Figure 8.4: Parent-child bow pattern

Pearl claims, without proof, that the presence of the bow-pattern prevents the identification of the causal effect $p(y \mid \tilde{x})$ from observational data “since any portion of the observed dependence between X and Y may always be attributed to spurious dependences mediated by U ”. This claim can be verified as follows.

Suppose that, by successive application of Rules 1–3, we are able to reduce $p(y \mid F_X = x)$ to a form involving only measured variables under the idle regime. Then at some point we shall have to have invoked either Rule 2, in the form $Y \perp_{\mathcal{D}_{\bar{Z}}} F_X \mid (X, Z, W)$, or Rule 3, in the form $Y \perp_{\mathcal{D}_{\bar{Z}}} F_X \mid (Z, W)$. But by Remark 8.2 and Remark 8.1, we shall never be able to find measured variables Z, W for which either of these graph-separation properties holds. Consequently we shall not be able to reduce $p(y \mid F_X = x)$ to an identifiable form by applying the rules of the *do*-calculus. Since these rules are complete, it follows that the existence of a bow-pattern between X and Y precludes any possibility of identifying the causal effect of X on Y .

Remark 8.3 If we are willing to make additional assumptions it may become possible to identify $p(y \mid F_X = x)$ in the above case: see Chapter 11.

Remark 8.4 It is clear from Example 8.1 that we can not remove the condition $Y \in \text{ch}(X)$.

Chapter 9

Confounding And Sufficient Covariates

We consider a response variable Y , binary treatment variable T , and associated treatment regime indicator F_T . We shall impose the following reasonable *positivity requirement*:

Condition 9.1 *In the observational regime $F_T = \emptyset$, both values of T occur with positive probability.*

We are fundamentally interested in comparing the distributions of Y given, respectively, $F_T = 1$ (active treatment) and $F_T = 0$ (control). Because these are interventional regimes, such a comparison can be interpreted causally, and in particular used to address a decision problem, as described in §2.4. A simple and often useful comparison is the *average causal effect*¹ of T on Y , defined as:

$$\text{ACE} := \text{E}(Y \mid F_T = 1) - \text{E}(Y \mid F_T = 0). \quad (9.1)$$

However, when we have data from an observational regime, rather than directly from the interventional regimes of interest, making such causal comparisons can be problematic.

9.1 Example: Normal regression

We will use the following concrete example to illustrate the more abstract concepts below.

Example 9.1 In addition to the treatment variable T and univariate response variable Y , we have a set $\mathbf{U} = (U_1, \dots, U_p)'$ of *covariates* measured on each unit.

The conditional distribution of Y given (T, \mathbf{U}) is the same in all regimes: specifically, it is normal, with mean

$$\text{E}(Y \mid T, \mathbf{U}, F_T) = d + \delta T + \mathbf{b}'\mathbf{U} \quad (9.2)$$

and variance σ^2 .

In the observational regime $F_T = \emptyset$, $T = 0$ or 1 , each with probability $\frac{1}{2}$; and the distribution of \mathbf{U} given $T = t$ ($t = 0, 1$) is multivariate normal with mean $\boldsymbol{\mu}_t$ and dispersion matrix Σ . In particular, the marginal distribution Q of \mathbf{U} is a 50-50 mixture of these two multivariate normal distributions, having mean $\bar{\boldsymbol{\mu}} := \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. We further suppose that U has this same marginal distribution Q in each of the two interventional regimes, $F_T = 0, 1$. We assume all the parameters of all the above distributions to be known.²

Note that $\text{E}(Y \mid F_T = t) = d + \delta t + \mathbf{b}'\bar{\boldsymbol{\mu}}$. So the ACE of (9.1) is just δ , which can thus be interpreted causally, and will be taken as the principal target of causal interest. \square

¹A better term might be “causal average effect”

²—except for §10.3 below.

9.2 No confounding

Consider the property:

$$Y \perp\!\!\!\perp F_T \mid T. \quad (9.3)$$

This asserts that the distribution of Y given T will be the same, whether the variables arise naturally or T is set by intervention. Since, for $t = 0, 1$, $F_T = t \Rightarrow T = t$, this is equivalent to requiring that, for $t = 0, 1$, the conditional observational distribution of Y given $T = t$ be the same as the marginal distribution of Y under an intervention that sets T at t .³ When this holds, causal enquiries about the ‘effect of T on Y ’, which we regard as relating to the distributions of Y given $F_T = t$ for various settings t , can be addressed directly from observational data in this situation: in particular, we can identify the average causal effect, since then

$$\text{ACE} = \text{E}(Y \mid T = 1, F_T = \emptyset) - \text{E}(Y \mid T = 0, F_T = \emptyset). \quad (9.4)$$

In such a case we say that there is *no confounding* (of the effect of T on Y).

Note that our definition (9.3) of ‘no confounding’ is absolute, rather than relative to a given model—as is the case, for example, for Definition 6.2.1 of Pearl (2000). But it can certainly be helpful to express this by means of a graphical model. The augmented DAG of Figure 9.1 describes the situation in which (9.3) applies.



Figure 9.1: No confounding

9.2.1 Potential responses

In the potential response framework, the “no confounding” condition is usually expressed as

$$T \perp\!\!\!\perp (Y_0, Y_1) \mid F_T = \emptyset, \quad (9.5)$$

which says: “In the observational regime, treatment T is independent of the pair (Y_0, Y_1) of potential responses.” In fact, it is sufficient to make the weaker assumption of independence for each potential response separately:

$$T \perp\!\!\!\perp Y_t \mid F_T = \emptyset \quad (t = 0, 1). \quad (9.6)$$

When (9.6) can be assumed, the observational distribution of $Y \equiv Y_T$ given $T = t$ is just the marginal distribution of Y_t , which, it is assumed, is the response to intervening with $F_T = t$. Thus (9.3) follows.

9.3 Confounding

As the examples of § 1.2 illustrate, (9.3) is a bold assumption that will often be inappropriate. In this case we say there is *confounding*.

Example 9.2 In Example 9.1, the expectation of Y given $T = t$ and F_T is $d + \delta t + \mathbf{b}'\boldsymbol{\mu}_t$ for $F_T = \emptyset$, but $d + \delta t + \mathbf{b}'\bar{\boldsymbol{\mu}}$ for $F_T = t$. So (9.3) does *not* hold in this case. \square

Example 9.3 Consider again Example 1.8. It appears from Tables 1.2 and 1.3 there that the variable “sex” is strongly associated with survival.

We can also examine the relationship, in the study, between sex and treatment (Table 9.1). This makes it apparent that treatment is strongly confounded with sex, so that, in the overall

³We need Condition 9.1 here to ensure that both these conditional observational distributions are well-defined.

	Male	Female	Total	% Male
New treatment	300	100	400	75%
Standard treatment	100	300	400	25%

Table 9.1: Confounding by sex

comparison in Table 1.1, we are not comparing like with like. In particular, given only these summary data, we do not know whether to ascribe the differences we see in the survival rates to differences in treatment, or to sex differences. The variable “sex”, being associated with both treatment and survival, is a *confounder* for the effect of treatment on survival. It would be rash to assume (9.3) and thereby treat the overall survival rates of Table 1.1 as predictive of what would happen to a new patient under either treatment.

Can we however trust the more detailed data of Tables 1.2 and 1.3? If we want to treat a new male patient, could we use the survival rates seen in Table 1.2 to predict his response under either treatment? This might or might not be appropriate—there might be further, unmeasured, variables that confound this interpretation too.

If we *can* assume that there is no further confounding, after adjusting for sex, we can say that sex is a *sufficient covariate* in this problem. \square

9.4 Sufficient covariate

Suppose we are happy to assume that, for a certain additional variable⁴ U , the joint structure of (F_T, T, U, Y) has the following two properties:

$$U \perp\!\!\!\perp F_T \quad (9.7)$$

$$Y \perp\!\!\!\perp F_T \mid (U, T). \quad (9.8)$$

Properties (9.7) and (9.8) can also be expressed by means of the DAG of Figure 9.2 (where the labels a and b should be ignored for the moment).

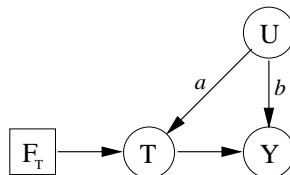


Figure 9.2: Sufficient covariate

A variable U for which conditions (9.7) and (9.8) hold is termed a *sufficient covariate* (for the effect of T on Y).⁵ In general there may be many choices (or none) for such a variable or set of variables.

Example 9.4 In Example 9.1, taking $U \equiv \mathbf{U}$, properties (9.7) and (9.8) are readily seen to hold. So \mathbf{U} is a sufficient covariate.⁶ \square

⁴We allow U to be multivariate, so constituting a set of variables rather than just one.

⁵This definition is more general than that of Lauritzen (2000), which involves additional unmeasured variables.

⁶Given that, in this example, the properties of (T, U) were introduced in terms of a marginal distribution for T and a conditional distribution for U given T , it may appear odd to represent it by Figure 9.2, which is naturally read as specifying a marginal distribution for U and a conditional distribution for T given U . However, it is important to realize that it is not the directions of the arrows in a DAG that carry causal meaning, but rather the implied conditional independences as can be read off using moralization. As we have constructed Example 9.1 we do indeed have (9.7) and (9.8), so that Figure 9.2 is indeed a valid representation.

Condition (9.8) is essentially the same as (9.3), but further conditioned on U . When this holds, if we could always observe U we could identify the conditional distribution of Y given T and U in the observational regime, and safely transfer this to the interventional regime. This would enable a new subject, with measured U , to address his treatment decision problem.

In particular, under Condition (9.8) we can use the observational data to identify the *specific causal effect* of T on Y , defined as follows:

Definition 9.1 The *specific causal effect* of treatment, for value u of U , is:

$$\text{SCE}_U(u) := \text{E}(Y \mid U = u, F_T = 1) - \text{E}(Y \mid U = u, F_T = 0). \quad (9.9)$$

Thus $\text{SCE}_U(u)$ is an average causal effect, but restricted to the subpopulation of individuals having the specified value u of U . We also denote by SCE_U the random variable (function of U) $\text{SCE}_U := \text{SCE}_U(U) = \text{E}(Y \mid U, F_T = 1) - \text{E}(Y \mid U, F_T = 0)$. \square

When (9.8) holds, and using $F_T = t \Rightarrow T = t$, $t = 0, 1$, we shall have

$$\text{SCE}_U := \text{E}(Y \mid U, T = 1, F_T = \emptyset) - \text{E}(Y \mid U, T = 0, F_T = \emptyset), \quad (9.10)$$

so that SCE_U is observationally identifiable.

Example 9.5 In Example 9.1, it is easily seen that SCE_U is the constant δ , in this special case being independent of the value u of U . \square

Condition (9.7) requires that the distribution of U be the same in all regimes: this is perhaps less fundamental than (9.8), but generally useful. As we shall see in §9.9, even when we are not in a situation of “no confounding”, so long as we can observe a sufficient covariate U along with T and Y we can still identify the desired interventional distributions for Y given $F_T = t$ from data gathered under the observational regime $F_T = \emptyset$. This is termed the case of *no unobserved confounders*.

Typically (though not necessarily), in order for condition (9.8) to be reasonable we would want U to comprise a large set of variables: intuitively, we want U to account for the totality of those individual characteristics that are relevant to the process whereby T generates Y . However, (9.7) acts as a constraint on just how much we can throw into U . In any case, whatever intuitions we may use to nominate a sufficient covariate U , the essential point is to ensure that we are satisfied as to the appropriateness of both (9.7) and (9.8).

9.5 Allocation process

Here we describe one natural way in which the properties described in §9.4 might come about.

We have a population of exchangeable subjects (see §4.1.4), to each of which we can assign a treatment, T .

Let \mathbf{U} be a set of “pre-treatment variables”. These might be genuine “covariates”, *i.e.* pre-existing subject characteristics, or other background variables. We can also include in \mathbf{U} variables generated by further randomization, with probabilities perhaps tuned to observed subject characteristics. We here use the term “covariate” widely, to include all such variables.

We now conceive of a binary *allocation indicator* D , itself in \mathbf{U} , describing which treatment would be chosen in the absence of any external intervention. This will typically be highly correlated with individual subject characteristics that might themselves be predictive of response Y , so generating confounding.

Let P_t ($t = 0, 1$) denote the joint distribution of all relevant variables (*viz.* \mathbf{U} and the response Y) when a subject is assigned treatment t (*i.e.* under interventional regime F_t).

We assume:

- (i). The distribution of \mathbf{U} is the same under both P_0 and P_1 .

- (ii). A subject's response to a treatment does not depend on whether or not the treatment has been externally imposed.⁷

Consider now an observational regime, F_{\emptyset} , constructed as follows. We select a subject at random from the population, observe his value of D , assign the corresponding treatment ($T = D$), and finally observe his response Y . (Note that T is then functionally determined by D and F_T : $T = t$ when $F_T = t$ ($t = 0, 1$), while $T = D$ when $F_T = \emptyset$.)

The observational joint distribution P_{\emptyset} can now be found. The covariates \mathbf{U} (which include D) have the same joint distribution as in either interventional regime (the same in both by assumption (i)); while the conditional distribution of Y given $\mathbf{U} = \mathbf{u}$ is the same as in regime F_d , where d is the value of D in \mathbf{u} .

It is easily seen that \mathbf{U} is a sufficient covariate. In fact, as will be shown in §10.2.1 below, the allocation indicator D is itself a sufficient covariate.

9.6 Potential responses

In our approach we regard a covariate as a quantity that could be fully measured, at any rate in appropriate circumstances. But if we take a purely formal approach we can also admit more general quantities.

In particular, suppose we are willing to conceive of the existence of the pair (Y_0, Y_1) of potential responses. These are supposed to have the same values, and *a fortiori* the same joint distribution, no matter what regime operates. That is:

$$(Y_0, Y_1) \perp\!\!\!\perp F_T. \quad (9.11)$$

Further, when the pair (Y_0, Y_1) and the treatment applied T are known, the actual response Y is fully determined: $Y = Y_T$, no matter what regime operates. This implies in particular:

$$Y \perp\!\!\!\perp F_T \mid (Y_0, Y_1, T). \quad (9.12)$$

Comparing (9.11) and (9.12) with (9.7) and (9.8), we see that, formally at least, we can always treat $U^* \equiv (Y_0, Y_1)$ as a sufficient covariate.

9.7 Confounding

It would be nice if from (9.7) and (9.8) we could deduce the “no confounding” property (9.3). However in general this will not be so. For, using the moralization criterion to query Figure 9.2 to check whether it implies $Y \perp\!\!\!\perp F_T \mid T$, we find we need to add a moralization link between U and F_T , so creating a path $Y-U-F_T$ from Y to F_T avoiding T .

A sufficient covariate U , satisfying (9.7) and (9.8), is also called *potential confounder*. It is a *non-confounder* in the very special case that (9.3) holds; otherwise a *confounder*.⁸ In general, when all potential confounders are unmeasured we can not identify causal effects from observational data.

9.8 Nonconfounding

When will a potential confounder in fact be a non-confounder, *i.e.* when can we deduce (9.3) from (9.7) and (9.8)? Two different sufficient conditions are given in the following.

⁷Assumption (ii) may well be unreasonable in some (especially economic or sociological) contexts, since a subject might behave differently if forced to take a treatment than he would if he himself had chosen to do so.

⁸Many workers would use this description for a single component of a multivariate sufficient covariate, not necessarily sufficient of itself. We shall not have any use for this extension.

Lemma 9.1 *Suppose that, in addition to (9.7) and (9.8), either of the following conditions holds:*

$$Y \perp\!\!\!\perp U \mid T \quad (9.13)$$

$$T \perp\!\!\!\perp U \mid F_T. \quad (9.14)$$

Then $Y \perp\!\!\!\perp F_T \mid T$.

Proof. These results can be shown directly by algebraic manipulation of conditional independence properties using P1–P5. Alternatively we can use graphical representations. Thus under (9.13) the arrow b in Figure 9.2 is absent, and we can represent the problem by Figure 9.3. The

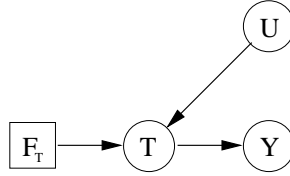


Figure 9.3: Irrelevance

property $Y \perp\!\!\!\perp F_T \mid T$ is then easily verified using the moralization criterion. If, instead, we assume (9.14), we can drop arrow a from Figure 9.2, thus obtaining Figure 9.4. We can again read off the desired property $Y \perp\!\!\!\perp F_T \mid T$. \square

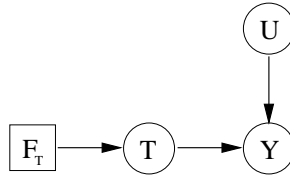


Figure 9.4: Randomization

Condition (9.13) says that the conditional distribution of response Y given treatment T and covariate U (which, by (9.8) has already been assumed unaffected by whether treatment arose naturally or by intervention), is not in fact affected by the value of U . It is indeed obvious then that we can entirely ignore U , *even if we were to observe it*; and that the conditional distribution of Y given treatment T (which is the same as that given both T and U) is unaffected by whether treatment arose naturally or by intervention. In this case we could call U *irrelevant* for Y . In a sense condition (9.13) represents a rather trivial way of obtaining (9.3), and in cases where it holds we might have been just as willing to assume (9.3) directly.

More interesting is (9.14). Because the conditional distribution of T given F_T is degenerate whenever $F_T \neq \emptyset$, this condition is only non-trivial for the case $F_T = \emptyset$. It is thus equivalent to requiring that T and U be independent when both arise naturally—in which case we can say that U is *unassociated* with treatment. Since this would occur in an experiment with completely randomized assignment of T (at any rate when U is a pre-treatment quantity), we may describe (9.14) as the ‘randomization condition’. Under this condition, even though U might not be irrelevant, and hence would be useful were it to be measured, when it is not available it does not bias the assessment of the distribution of response given only treatment, whose observational and interventional versions will still be identical.

9.8.1 Other conditions

It may be possible to demonstrate (9.3) under other conditions than those considered above. A simple example is given in Figure 9.5, where $U = (U_1, U_2)$: we readily read off, by moralization, the “no confounding” property $Y \perp\!\!\!\perp F_T \mid T$, even though neither (9.13) nor (9.14) need hold.

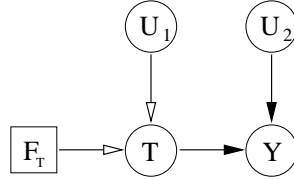


Figure 9.5: Nonconfounding

More generally, given a model involving specified relations between Y , T , F_T and some additional variables, it will sometimes be possible to demonstrate (9.3), either directly, through manipulation of conditional independence properties, or indirectly, by applying the moralization criterion to a suitable graphical representation.

9.8.2 ‘No unobserved confounders’

In general, in order to be able to proceed without taking any account of confounding, it is necessary that, for *any* potential confounder U , it in fact turns out to be a non-confounder—the situation described by the phrase ‘no unobserved confounders’. Now this initially might appear to require explicit consideration of all possible potential confounders—a task that would generally be impossible in practice (and perhaps even in principle). Fortunately this is not required. So long as we can identify *just one* potential confounder U that is a non-confounder, we can deduce (9.3), and hence will be justified in interpreting the observational distributions of Y given T as equally meaningful under intervention. In particular, it will be enough to have (9.13) or (9.14) hold for *some* potential confounder. Then any other potential confounder must also be a non-confounder (even though it might not satisfy either (9.13) or (9.14)—since while these conditions are sufficient, they are not necessary).

9.9 Deconfounding

Suppose that we have observational data on a sufficient covariate U , as well as on Y and T . However, we can not observe any covariate information for the new subject S for whom a treatment decision is required. For that purpose we would thus want to identify and compare the ‘marginal’ interventional distributions $p(y \mid F_t = t)$.

Now (9.7) and (9.8) are exactly the conditions required for the application of the “back-door” formula of Theorem 8.1. Consequently, we can calculate:

$$p(y \mid F_T = t) = \sum_u p(y \mid F_T = t, u) p(u \mid F_T = t) \quad (9.15)$$

$$= \sum_u p(y \mid T = t, u) p(u). \quad (9.16)$$

Thus in the presence of (9.7) and (9.8), if we can observe U we can identify the ‘marginal causal effect’, on Y , of setting T to t , from an observational study in which (Y, T, U) are all recorded.

9.9.1 Complete confounding

To be more precise, for deconfounding using (9.16) we also require an additional ‘conditional positivity condition’: under the observational regime $F_T = \emptyset$, for each $t = 0, 1$ we want $\Pr(T = t \mid$

$U = u) > 0$ whenever $\Pr(U = u) > 0$. Otherwise we will not be able to identify $p(y | T = t, u)$.

For example, when our sufficient covariate is the allocation indicator D , (9.16) would give $p(y | F_T = 1) = p(y | T = 1, D = 0) \Pr(D = 0) + p(y | T = 1, D = 1) \Pr(D = 0)$. But we can not identify $p(y | T = 1, D = 0)$, since the combination $T = 1, D = 0$ does not occur in the observational regime.

9.9.2 External standardization

In some cases, we might be willing to assume (9.8), but not (9.7), relying instead on other, external, information to identify $p(u | F_T = t)$ in (9.15). For example, if U is sex, and we have (by design or accident) sampled different numbers of men and women, yielding say $\Pr(M | F_T = \emptyset) = 0.8$, we would not want to apply this observational sex ratio to a new subject. If we know that the new subject is male, we should use $\Pr(M | F_T = t) = 1$ in (9.15). If we can not observe the subject's sex, we might reasonably take $\Pr(M | F_T = t) = 0.5$.

9.10 Average causal effect

As a simple consequence of (9.16), we find, for any sufficient covariate U :

$$\text{ACE} = \text{E}\{\text{E}(Y | T = 1, U)\} - \text{E}\{\text{E}(Y | T = 0, U)\} \quad (9.17)$$

$$= \text{E}(\text{SCE}_U) \quad (9.18)$$

where ACE is the average causal effect given by (9.1), and SCE_U the specific causal effect, observationally identifiable by (9.10). As in §9.9.2, the observational marginal distribution of U used to calculate the expectation in (9.18) might be replaced by some externally justified distribution when we can not assume (9.7).

9.10.1 Potential responses

In the potential response framework, we can formally take $U = (Y_0, Y_1)$. Then we have $\text{E}(Y | T = t, U) = Y_t$, so that

$$\text{ACE} = \text{E}(Y_1 - Y_0), \quad (9.19)$$

the expectation of the *individual causal effect* $\text{ICE} := Y_1 - Y_0$. (Note however that ICE is always unobservable, since we can not observe both Y_0 and Y_1 simultaneously.)

Chapter 10

Reduction Of Sufficient Covariate

Let U be a sufficient covariate. It is often possible to reduce the information in the variable U by replacing it by some $V \preceq U$ while retaining the sufficiency property. Note that the condition

$$V \perp\!\!\!\perp F_T \tag{10.1}$$

is automatically satisfied, by P3. So for any such putative V we only need check:

$$Y \perp\!\!\!\perp F_T \mid (V, T). \tag{10.2}$$

There are two main ways of proceeding, corresponding to the two arrows b and a in Figure 9.2 describing, respectively, the effect of U on Y and on T .

10.1 Reduction of effect on Y

By Condition 9.1, for $t = 0, 1$, each of $\Pr(Y \in A \mid U, T = t, F_T = t)$, $\Pr(Y \in A \mid U, T = t, F_T = \emptyset)$ is defined, as a function of U , up to a set of values having probability 0 under the distribution of U (which by (9.7) is the same in all regimes). Then (9.8) asserts that these are (almost surely) identical. Thus if we define $\Pr(Y \in A \mid U, T = t) := \Pr(Y \in A \mid U, F_T = t)$, this will serve as a conditional distribution for Y given $(U, T = t)$ in all regimes.

We shall show that (10.2) holds if the following condition is satisfied:

Condition 10.1 (Response-sufficient reduction) For $t = 0, 1$, $Y \perp\!\!\!\perp U \mid (V, F_T = t)$.

This says that, for each treatment decision, once V is known further information about U is of no value for predicting Y .

Because of Condition 9.1, Condition 10.1 is equivalent to saying that, for each $t = 0, 1$, the observational conditional distribution $\Pr(Y \in A \mid U, T = t, F_T = \emptyset)$ —a distribution that is in fact common to all regimes, on account of (9.8) as seen above—in fact depends on U only through its reduction V : a property that can be tested using observational data. It can also be expressed as:

$$Y \perp\!\!\!\perp (U, F_T) \mid (V, T). \tag{10.3}$$

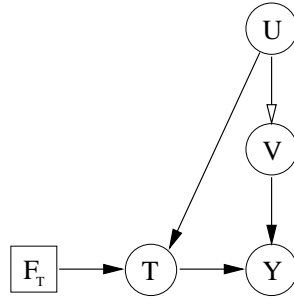
(Note that (10.3) already incorporates (9.8), using P4 and the property $V \preceq U$). Property (10.2) now follows, showing that V is a sufficient covariate.

Further, for any covariate X that is a function of U , we can easily show that:

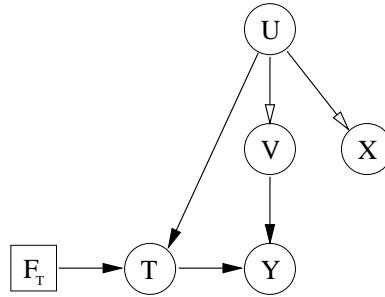
$$V \perp\!\!\!\perp F_T \mid X \tag{10.4}$$

$$Y \perp\!\!\!\perp F_T \mid (X, V, T). \tag{10.5}$$

That is to say, V remains a sufficient reduction of U even after we condition throughout on X .

Figure 10.1: Reduction of sufficient covariate between U and Y

The conditional independence properties (9.7) and (10.3) can be represented by the DAG of Figure 10.1 (the omission of an arrow from V to T , representing the property $V \perp\!\!\!\perp T \mid U$, is justified by the fact that, given U , V is non-random and hence independent of anything). Their implications (10.1) and (10.2) are easily read off using the moralization criterion. Similarly (10.4) and (10.5) can be read off the extended DAG of Figure 10.2, where again the functional dependence of X on U justifies representing X as conditionally independent of all other variables, given U .

Figure 10.2: Conditional sufficiency of V , given X

Example 10.1 In Example 9.1, a linear transformation $V = AU$ is a response-sufficient reduction of U whenever b is in the range-space of A' . \square

10.1.1 Minimal response-sufficiency

Suppose that we have a collection $\mathcal{V} = \{V_\alpha : \alpha \in \mathcal{A}\}$ of response-sufficient reductions of U . Then, for each $t = 0, 1$, the conditional distribution of Y given U , consequent on treatment decision $F_T = t$, actually depends on U through V_α alone—for any $\alpha \in \mathcal{A}$. Intuitively,¹ this means that its actual dependence on U can only be through whatever information is common to all the V_α 's—which could itself be embodied in a variable $V_0 \preceq V_\alpha$ (all α). Then V_0 likewise would satisfy Condition 10.1 and hence be a sufficient covariate. Taking \mathcal{V} to be the set of *all* response-sufficient reductions of U , V_0 is thus the *minimal response-sufficient* reduction of U .

Example 10.2 In Example 9.1, the minimal response-sufficient reduction of U is $V_0 := b'U$. \square

¹There are some technical difficulties in making this argument fully rigorous, very similar to those involved in the general construction of a minimal sufficient statistics for parametric inference (Dawid 1980). However these are of no importance for most practical applications.

Caution: The above argument does not work if we can not assume all the V_α are response-sufficient reductions of the same variable U which is already sufficient. In particular, if we only know that V_1 and V_2 are sufficient covariates, it may not be possible to find a common sufficient reduction V_0 of them both.

10.2 Reduction of effect on T

Consider now the alternative condition, for $W \preceq U$:

Condition 10.2 (Treatment-sufficient reduction) $T \perp\!\!\!\perp U \mid (W, F_T = \emptyset)$.

This says that, in the observational regime, treatment assignment T depends on U only through its reduction W —a property that, like Condition 10.1, can be tested using observational data. An equivalent characterization is that, for the (two-member) family of observational distributions for “data” U , given “parameter” T , the “statistic” W is a sufficient statistic.

Because T is in any case non-random in the interventional regimes $F_T = 0$ and $F_T = 1$, Condition 10.2 is equivalent to

$$T \perp\!\!\!\perp U \mid (W, F_T). \quad (10.6)$$

We can represent the conditional independence relations (9.7), (9.8) and (10.6) by means of the DAG of Figure 10.3.² Now $Y \perp\!\!\!\perp F_T \mid (W, T)$ follows on applying the moralization criterion. Hence W is a sufficient covariate.

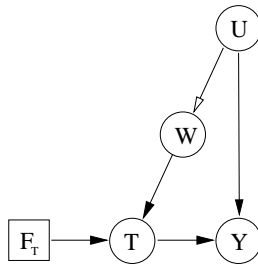


Figure 10.3: Reduction of sufficient covariate between U and T

Again, W will remain sufficient after conditioning on a further covariate $X \preceq U$, as can be read off from Figure 10.4.

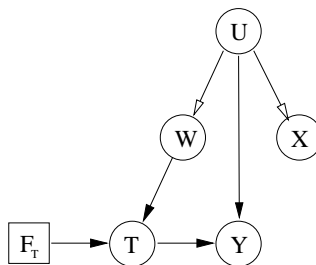


Figure 10.4: Conditional sufficiency of W , given X

²Again, we do not need an arrow from W to Y because $W \preceq U$.

Example 10.3 In the observational regime of Example 9.1,

$$W \equiv \mathbf{c}'\mathbf{U} \quad (10.7)$$

with

$$\mathbf{c} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad (10.8)$$

is the *Fisher linear discriminant function* (LDF) for separating the two multivariate normal distributions for \mathbf{U} , given T . It is well known, and straightforward to check, that $\mathbf{U} \perp\!\!\!\perp T \mid W$, so that W is a treatment-sufficient reduction of \mathbf{U} —as is any linear transformation $\mathbf{W} = \mathbf{C}\mathbf{U}$ such that \mathbf{c} is in the range-space of \mathbf{C}' . \square

10.2.1 Allocation process

Consider again the allocation process of §9.5, with allocation indicator D . We suppose we start with some sufficient covariate $U \in \mathbf{U}$.

Since, in the observational regime, $T \equiv D$, Condition 10.2 is equivalent to the requirement that (in the joint distribution of covariates, which is common to all regimes):

$$D \perp\!\!\!\perp U \mid W. \quad (10.9)$$

In particular, (10.9) holds if we take $U = \mathbf{U}$, $W = D$, so the allocation indicator D itself is a sufficient covariate—as similarly is any subset of \mathbf{U} containing D . Alternatively, (10.9) will hold if D is generated by using a randomizing device with probabilities depending on U only through the value of W .

It may appear surprising that (10.9) does not explicitly include a requirement that the additional information in D , over and above that in W , be of no value in predicting the response Y . However, Theorem 10.1 below shows that, so long at any rate as we are comparing only two treatments, this property is already implicit in (10.9).

Theorem 10.1 *In the setting of §9.5, with binary treatment variable T , response variable Y , and allocation indicator D , suppose W is a sufficient covariate. Then for $t = 0, 1$*

$$Y \perp\!\!\!\perp D \mid (W, F_T = t). \quad (10.10)$$

Proof. From the comment after (10.9), (W, D) is sufficient, so that

$$Y \perp\!\!\!\perp F_T \mid (W, T, D). \quad (10.11)$$

If, in addition, W alone is to be sufficient, we must also have:

$$Y \perp\!\!\!\perp F_T \mid (W, T). \quad (10.12)$$

According to (10.11) and (10.12), conditional on any fixed values (w, t) for (W, T) , Y must be independent of F_T both marginally and conditionally on D . But (Yule 1903; Dawid 1980) since D is binary these two properties can only hold if at least one of the following holds:

$$Y \perp\!\!\!\perp D \mid (W = w, T = t, F_T) \quad (10.13)$$

$$D \perp\!\!\!\perp F_T \mid (W = w, T = t). \quad (10.14)$$

Now (10.13) holds trivially for $F_T = \emptyset$, so only has real bite for $F_T = 0$ or 1 . In these cases it is equivalent to:

$$Y \perp\!\!\!\perp D \mid (W = w, F_T = t). \quad (10.15)$$

If, on the other hand, (10.14) holds for some (w_0, t_0) , we have

$$\Pr(D = t_0 \mid W = w_0, T = t_0, F_T = \emptyset) = \Pr(D = t_0 \mid W = w_0, F_T = t_0). \quad (10.16)$$

The left-hand side of (10.16) is 1. The right-hand side is $\Pr(D = t_0 \mid W = w_0, F_T)$ for any regime F_T , since (W, D) is a covariate and thus independent of F_T . So if (10.14) holds for (w_0, t_0) , then conditional on $W = w_0$, D is almost surely constant, in any regime. But this itself implies (10.15) (for either value of t). Thus (10.15) holds for all (w, t) , so showing (10.10). \square

10.2.2 Minimal treatment-sufficiency

Suppose that we have a collection $\mathcal{W} = \{W_\alpha : \alpha \in \mathcal{A}\}$ of treatment-sufficient reductions of U , all satisfying Condition 10.2. Then the observational conditional distribution of Y given U actually depends on U through W_α alone, for any $\alpha \in \mathcal{A}$. Arguing as in the response-sufficient case, we can construct a *minimal treatment-sufficient* reduction W_0 of U . This is in fact the same as the definition and construction for determining the *minimal sufficient statistic* for the observational distributions of “data” U given “parameter” T . One choice for this is the *propensity score* $\Pi := \Pr(T = 1 \mid U)$.

Example 10.4 In Example 10.3, the LDF variable W is a minimal treatment-sufficient covariate. The propensity score $\Pi = \Pr(T = 1 \mid U)$ is an increasing function of W , given by:

$$\text{logit } \Pi = W - m \tag{10.17}$$

where

$$\begin{aligned} m &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \Sigma^{-1} \bar{\boldsymbol{\mu}} \\ &= \text{E}(W \mid F_T = \emptyset). \end{aligned}$$

□

10.3 Propensity scoring in practice

The *propensity score* $\Pi := \Pr(T = 1 \mid U)$ was introduced in § 10.2.2 as a minimal treatment-sufficient reduction of a sufficient covariate U for a binary treatment T . Rubin (2007) argues that the analysis of observational data is better if based on appropriate adjustment for the propensity score Π , rather than for a response-sufficient covariate V . One practical point in favour is that, because the propensity score Π can be identified without even looking at the outcome variable, it is more “objective” and so guards against deliberate or subconscious bias. Propensity adjustment can also be considered as a way of mimicking (conditionally) random treatment assignment: all subjects with the same propensity score have, independently, the identical probability of being assigned to active treatment (or to control).

All the above is fine when we know the true propensity score Π , or (essentially equivalent) a minimal treatment-sufficient covariate W . In practice, however, we will not know the underlying probability structure, and so will have to use the observational data to estimate the form of W , *e.g.* by fitting a model for the dependence of T on U . (In any case we have to assume that we are starting from a sufficient covariate U , which it is usually impossible to be sure of). When the dimension of U is not small compared with the number of data-points, there is every danger that we will “overfit” to random patterns of the data in hand, leading to poor estimates. There have been numerous claims that this is not a real problem for propensity scoring: it is even suggested that adjusting for an estimated propensity score is superior to using the true score. However, this can not be generally true, as the following example shows.

Example 10.5 Consider Example 9.1, with minimal treatment-sufficient covariate W given by (10.7) and (10.8). Now from (9.2)

$$\text{E}(Y \mid T, W, F_T) = d + \delta T + \mathbf{b}'\text{E}(\mathbf{U} \mid W), \tag{10.18}$$

where the final term does not involve T on account of (10.6), and is a linear function of W . Consequently (10.18) has the form:

$$\text{E}(Y \mid T, W) = \alpha + \delta T + \beta W. \tag{10.19}$$

It follows that the coefficient of T in the regression of Y on T and W is exactly the same as in the regression on T and all of \mathbf{U} , namely the ACE, δ .

This appears to be good news—we can reduce the multivariate covariate \mathbf{U} to the univariate covariate W without affecting the interpretation of the regression coefficient of T as a causal effect.

But consider now what happens when we do not know the parameters of the model but have to estimate them from data. The standard procedure would estimate $\boldsymbol{\mu}_i$ by the relevant sample mean $\bar{\mathbf{x}}_i$, and (up to an unimportant scale-factor) Σ by the within-group sum-of-squares-and-products matrix S . After substituting these estimates for the population parameters, application of the identical manipulations that yield the population LDF in the known-parameter case will now yield the corresponding sample LDF.

Suppose we proceed by first identifying this sample LDF, say \widehat{W} (a linear combination of the (U_i) with coefficients estimated from the data), and then calculating the sample regression of Y on T and \widehat{W} . We would find, in analogy to (10.6) (applied now at the level of first- and second-order moments only) that the sample (estimated) regression of U on T and \widehat{W} does not involve T —in this sense, adjusting for \widehat{W} has brought the two sample distributions of U given T into agreement. In much of the propensity literature, this “balancing” behaviour is regarded as highly desirable, and considered as leading to improved accuracy for estimating the causal effect on the response. By contrast, even if we knew the true LDF W it would *not* in general possess this property in the sample. So it might appear, from examining the data, that adjusting for an estimated propensity-type variable would be *better* than adjusting for its true population counterpart.

Now the sample version of (10.19) shows that the coefficient $\widehat{\delta}$ of T in the sample regression of Y on T and the single variable \widehat{W} will be identical with that in the sample regression on T and the full multivariate covariate U . Since the estimate is unchanged, so necessarily is its precision. In particular, when the sample size n is not much greater than the dimension p of U , the variance of the estimator $\widehat{\delta}$ can be very large; moreover, this is entirely unaffected by whether we have calculated it by regressing Y directly on (p -variate) U , or indirectly by first forming \widehat{W} and then regressing Y on T and (univariate) \widehat{W} . Conducting a propensity analysis has had absolutely no effect on the estimate and its uncertainty.

In effect, the data-dredging that leads to \widehat{W} being a generally poor substitute for W counterbalances the increase in precision from reducing the covariates from p to 1—notwithstanding any seductive message in the sample that adjusting for \widehat{W} has brought the two sampling distributions of U given T into close alignment. \square

In practice, we will rarely be in a position to make strong distributional assumptions, such as multivariate normality, about the joint observational distribution of T and U . A common procedure would be to first fit a logistic regression model for the dependence of $\Pi = \Pr(T = 1 | U)$ on U , and then form groups of subjects whose estimated value for Π all fall in some interval, *e.g.* one of the intervals $[0, .2)$, $[\.2, .4)$, $[\.4, .6)$, $[\.6, .8)$, $[\.8, 1]$. Within any group, all subjects should have approximately the same propensity score, so that an observational comparison of the two treatments should mimic an experimental comparison. (Some of the groups, particularly at the extremes, might have very few subjects receiving one of the treatments: these might simply be discarded). Finally we can combine such comparisons across the groups.

Although the specific analysis of Example 10.5 will not apply to such more general analyses, the same sort of behaviour can be expected.

10.4 A complication

Although each of Condition 10.1 and Condition 10.2 is a sufficient condition for a reduction W of U to be a sufficient covariate, neither is necessary.

Example 10.6 Randomization. Suppose that, in the observational regime, we in fact have completely randomized allocation. That is, the binary allocation variable D is generated by a fair coin-flip, independently of all *pre-existing* covariates \mathbf{W} . In order to satisfy the previous requirement $D \in U$, we take the full set of “covariates” to be $U = (\mathbf{W}, D)$.

We assume the mere action of flipping and observing the coin can have no effect on the response, which thus depends only on the pre-existing covariates \mathbf{W} and the actually applied treatment T :

$$Y \perp\!\!\!\perp (D, F_T) \mid (\mathbf{W}, T). \quad (10.20)$$

We shall however suppose that \mathbf{W} affects response: *i.e.* for at least one value $t = 0$ or 1 ,

$$Y \not\perp\!\!\!\perp \mathbf{W} \mid T = t. \quad (10.21)$$

The assumed conditional independence properties can be represented graphically by Figure 10.5, a special case of Figure 9.5, so that (unsurprisingly) we have the “no confounding” property: $Y \perp\!\!\!\perp F_T \mid T$. In particular, this means that the trivial variable $\mathbf{0}$, identically equal to 0, is sufficient—and thus a sufficient reduction of U .

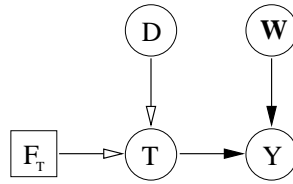


Figure 10.5: Randomization

However, on account of (10.21), $W = \mathbf{0}$ does *not* satisfy Condition 10.1. Further, $W = \mathbf{0}$ does *not* satisfy Condition 10.2. For in the observational regime we can not have $T \perp\!\!\!\perp U$, since $T \equiv D$ is completely determined by U (and is not constant).

Note that in this case application of the construction of § 10.2.2 would deliver D —the actual outcome of the coin-flip—as the “minimal treatment-sufficient reduction” of U (this would also be the associated propensity score), without giving any indication that it could be further reduced to $\mathbf{0}$ (so generating the constant propensity score 0.5) while retaining sufficiency. \square

10.5 Joint reduction?

Suppose V and W are, respectively, minimal response-sufficient and minimal treatment-sufficient reductions of U . Since $W \preceq U$, it follows from (10.3), on applying P4 and P3, that $Y \perp\!\!\!\perp F_T \mid (V, W, T)$. Hence $U^* := (V, W)$ constitutes a sufficient covariate, and moreover each of V, W is a sufficient reduction of U^* . It is tempting to conclude, by analogy with the arguments of §§ 10.1.1 and 10.2.2, that the information, U_0 say, in common to V and W will itself be a sufficient covariate. However this is typically not so (the **Caution** of § 10.1.1 applies): indeed, it will commonly be the case that U_0 is entirely vacuous, and so (in general) not sufficient.

Chapter 11

Instrumental Variables

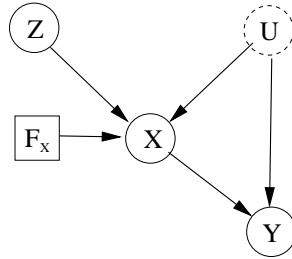


Figure 11.1: Instrumental variable

Suppose we would like to identify, if possible, the causal effect of a variable X on a response Y . Suppose further that we can describe the problem by means of Figure 11.1.

Here F_X is the usual intervention indicator for X , and U is an unobserved sufficient covariate for the effect of X on Y , so that

$$U \perp\!\!\!\perp F_X \tag{11.1}$$

$$Y \perp\!\!\!\perp F_X \mid (X, U). \tag{11.2}$$

In this case, however, we also have an additional observable variable Z (which in fact could be either a random variable or a decision variable—all our arguments will be conditional on Z) such that the following conditional independence properties (stronger than (11.1) and (11.2)) represented in Figure 11.1 hold:¹

$$U \perp\!\!\!\perp (Z, F_X) \tag{11.3}$$

$$Y \perp\!\!\!\perp (Z, F_X) \mid (X, U). \tag{11.4}$$

A variable Z satisfying (11.3) and (11.4) is termed an *instrumental variable* or *instrument* (for the effect of X on Y). The idea is that, in the observational regime, Z acts something like the intervention variable F_X to affect the value taken by X , without directly affecting the way Y depends on X . A good instrument is one which, in addition, is strongly associated with X . (An “instrument” that is independent of X is of no use whatsoever).

For the case of binary X , we sometimes have an *availability indicator*, a binary variable Z such that

$$Z = 0 \Rightarrow X = 0. \tag{11.5}$$

¹Figure 11.1 also encodes $Z \perp\!\!\!\perp F_X$, but since we will be arguing conditionally on Z this is of no significance.

Thus a subject having $Z = 0$ is simply not able to access the active treatment. When such Z is also an instrumental variable we may term it an *availability instrument*.

Note that, under an intervention at X ($F_X = x \neq \emptyset$), the arrows into X from Z and U are deleted, and we infer:

$$Y \perp\!\!\!\perp Z \mid F_X = x. \quad (11.6)$$

That is, the value of Z should not affect the behaviour of Y in response to an intervention at X . If this property can not be assumed, we can not regard Z as an instrumental variable. But there are other constraints too (although they involve unmeasured variables): thus (11.4) requires that the same variable U that “screens off” the effect of X on Y from the regime indicator F_X will also screen it off from Z in the observational regime.

Example 11.1 Randomized encouragement trial. In an encouragement trial patients are randomly assigned to take either active treatment or placebo, but might not comply with their assignment, so that the treatment actually taken, X , might differ from that assigned, Z . (Often, but not invariably, Z will be an availability instrument). The possible common dependence of X and Y on further unobserved patient characteristics will typically imply that Y can not be assumed independent of Z given X . However, because Z is randomized it should be independent of any such characteristics, and hence be an instrumental variable. \square

Example 11.2 Trans fatty acids. Trans fats, which are mainly found in partially hydrogenated vegetable oil and are common ingredients in thousands of food products, have been linked to raised blood cholesterol levels and heart disease.

Suppose we want to investigate the effect of trans fats on health outcomes. We could make use of the fact that in 2003 Denmark banned the use of trans fatty acids in packaged foods to treat living in Denmark as an availability instrument. For this analysis to be appropriate we have to assume (condition (11.3)) that all Nordic people are essentially exchangeable on dietary preference and other determinants of health. This assumption of similar diets and lifestyles is of course debatable.

In Britain, Sainsbury’s, Marks and Spencer and Asda have recently eliminated trans fatty acids from their own-brand foods; Tesco and Waitrose are planning to do so too. Thus which supermarket a consumer shops at acts as an availability variable. However, it is not likely to be an instrument, since it could well be associated with health-related life-style characteristics U . \square

Example 11.3 Non-randomized treatments. Suppose (McClellan *et al.* 1994) we are interested in the effect of invasive emergency procedures on patients suffering an acute myocardial infarction. Patients who receive different treatments tend to differ in both observable and unobservable health characteristics, so biasing naïve estimates of treatment effects. However the variable “distance to nearest emergency centre” is strongly associated with the type of treatment received, but can reasonably be considered independent of any patient characteristics that might affect response to treatment. It could thus be used as an instrument.

Examples such as this, where even though there is no experimental assignment of treatment there are nevertheless convincing reasons to accept the appropriateness of the assumptions (*e.g.* (11.3) and (11.4)) underlying the causal analysis, are sometimes termed “natural experiments”. \square

Example 11.4 Mendelian randomization. This is a type of natural experiment that currently looks very promising.

In the mid-1980s there was some observational evidence that low serum cholesterol level, X , could increase the risk of cancer, Y . However, account has to be taken of possible confounding. For example, a hidden tumour might induce a lowering of cholesterol in a future cancer patient, while factors such as diet and smoking might affect both cholesterol levels and cancer.

The gene APOE, U , is known to be associated with serum cholesterol, carriers of the E2 allele having particularly low levels. The value of U can reasonably be considered unrelated to socioeconomic position, lifestyle, and other such potential confounders. Moreover, lower cholesterol

levels in E2 carriers are present from birth, so can not be caused by pre-existing tumours. So U can serve as an instrumental variable for the effect of X on Y .

If low serum cholesterol level is really a risk factor for cancer, then patients should have more E2 alleles and controls should have more E3 and E4 alleles. Otherwise, APOE alleles should be equally distributed across both groups. \square

11.1 Causal inference

From § 8.3.2 we know that there can be no general expression for the desired causal effect of X on Y in terms of distributions identifiable from observational data on (Z, X, Y) . But if we are willing to make some stronger assumptions, or come away with weaker conclusions, we can make some progress.

11.2 Null hypothesis

The “causal null hypothesis” is $Y \perp\!\!\!\perp X \mid F_X \neq \emptyset$. A sufficient (though not necessary) condition for this is $Y \perp\!\!\!\perp X \mid U$, represented by the absence of the arrow $X \rightarrow Y$ in Figure 11.1. In that case we would also have $Y \perp\!\!\!\perp Z \mid F_X$ —a property that, in view of (11.6), only bites when $F_X = \emptyset$. But this is testable from observational data on (Z, Y) . If we discover it fails, we must deduce $Y \not\perp\!\!\!\perp X \mid U$. While it remains logically possible that the causal null hypothesis holds, this would only be by dint of perfectly counterbalancing values of the probabilities in the problem, so in realistic terms we can usually rule it out. That is, if we find the instrument is observationally associated with the response, we can deduce that X has a causal effect on Y .

11.3 Linear model

Suppose all the variables are univariate, and we can describe the dependence of Y on (X, U) in (11.4) by the *linear model*:

$$E(Y \mid X, U) = W + \beta X \quad (11.7)$$

for some function W of U .

Because (11.7) remains valid under an intervention $F_X = x$, we deduce

$$E(Y \mid F_X = x) = w_0 + \beta x$$

where $w_0 := E(W \mid F_X = x)$ is a constant independent of x , by (11.1). Hence β can be interpreted causally, as describing how the mean of Y changes in response to changes in the way we manipulate X . Our aim is to identify β .

Now by (11.4), (11.7) is also $E(Y \mid X, Z, U, F_T = \emptyset)$. So $E(Y \mid Z, F_T = \emptyset) = E(W \mid Z, F_T = \emptyset) + \beta E(X \mid Z, F_T = \emptyset)$. But by (11.3) the first term on the right-hand side term is constant. That is,

$$E(Y \mid Z = z, F_T = \emptyset) = \text{const.} + \beta E(X \mid Z = z, F_T = \emptyset). \quad (11.8)$$

Equation (11.8) relates two functions of z , both of which can be identified from observational data.² Consequently (so long as $E(X \mid Z = z, F_T = \emptyset)$ is not independent of z) we can identify the causal parameter β from such data.

Typically both $E(Y \mid Z, F_T = \emptyset)$ and $E(X \mid Z, F_T = \emptyset)$ would also be assumed linear in Z , and estimated by appropriate regression analyses; the required causal parameter β is then estimated by the ratio of the coefficients of Z in these two regressions.

²Indeed, we do not even need to observe all the variables (Z, X, Y) simultaneously: it is enough to have observational data on the pair (Z, X) and perhaps quite separate observational data on the pair (Z, Y) .

Note that we have not needed the full force of (11.3) in the above, but only:

$$\begin{array}{l} U \perp\!\!\!\perp F_X \mid F_X \neq \emptyset \\ U \perp\!\!\!\perp Z \mid F_X = \emptyset. \end{array}$$

11.4 Binary case

Now suppose that all the observable variables Z, X, Y are binary. Although we can not fully identify the ‘‘causal probability’’ $\omega_x := \Pr(Y = 1 \mid F_X = x)$ from observational data, we can develop inequalities it must satisfy.

11.4.1 Instrumental inequalities

We first note that, even leaving the variable U unspecified, the property of being an instrumental variable puts constraints on the observational distribution of (X, Y) given Z . For defining

$$\phi_{yx.z} := \Pr(Y = y, X = x \mid Z = z)$$

(where all probabilities are calculated under the observational regime) we have

$$\phi_{yx.z} = \mathbb{E}\{\Pr(X = x \mid Z = z, U) \Pr(Y = y \mid X = x, U)\}, \quad (11.9)$$

when the expectation is over the distribution of U . Now suppose we find $\phi_{00.0} = 1$. Then from (11.9) $\Pr(Y = 0 \mid X = 0, U) = 1$ almost surely, whence $\Pr(Y = 1 \mid X = 0, U) = 0$ almost surely. Again applying (11.9), this in turn implies we must have $\phi_{10.1} = 0$.

In fact it can be shown (Pearl 1995b) that a necessary set of conditions for the joint observational distribution to be compatible with having arisen from a model represented by Figure 11.1 is:

$$\phi_{0x.z_0} + \phi_{1x.z_1} \leq 1, \quad (11.10)$$

for all $x, z_0, z_1 \in \{0, 1\}$. Hence if we find that any of these inequalities is violated, we know that Z can not be an instrumental variable. Conversely, if all these inequalities hold it is formally possible to represent the joint distribution as arising from the model of Figure 11.1; although this does not mean that there need exist a real variable U for which it applies.

11.4.2 Causal inequalities

By somewhat sophisticated convex analysis, it can be shown (Balke and Pearl 1997; Dawid 2003) that, when Z is an instrumental variable, the causal probabilities ω_0, ω_1 can be bounded in terms of the observationally identifiable probabilities $(\phi_{yx.z})$, as follows:

$$\omega_0 \leq \min \left\{ \begin{array}{c} 1 - \phi_{00.0} \\ 1 - \phi_{00.1} \\ \phi_{01.0} + \phi_{10.0} + \phi_{10.1} + \phi_{11.1} \\ \phi_{10.0} + \phi_{11.0} + \phi_{01.1} + \phi_{10.1} \end{array} \right\} \quad (11.11)$$

$$\omega_0 \geq \max \left\{ \begin{array}{c} \phi_{10.1} \\ \phi_{10.0} \\ \phi_{10.0} + \phi_{11.0} - \phi_{00.1} - \phi_{11.1} \\ -\phi_{00.0} - \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{array} \right\} \quad (11.12)$$

$$\omega_1 \leq \min \left\{ \begin{array}{c} 1 - \phi_{01.1} \\ 1 - \phi_{01.0} \\ \phi_{10.0} + \phi_{11.0} + \phi_{00.1} + \phi_{11.1} \\ \phi_{00.0} + \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{array} \right\} \quad (11.13)$$

$$\omega_1 \geq \max \left\{ \begin{array}{c} \phi_{11.1} \\ \phi_{11.0} \\ -\phi_{01.0} - \phi_{10.0} + \phi_{10.1} + \phi_{11.1} \\ \phi_{10.0} + \phi_{11.0} - \phi_{01.1} - \phi_{10.1} \end{array} \right\}. \quad (11.14)$$

We further find that, for given ϕ , ω_0 and ω_1 can vary independently, subject only to the above inequalities. In particular (Balke and Pearl 1994), the bounds for the ‘‘causal effect’’ $\alpha := \omega_1 - \omega_0$ are given by: $\alpha_* \leq \alpha \leq \alpha^*$, where

$$\alpha_* := \max \left\{ \begin{array}{c} \phi_{00.0} + \phi_{11.1} - 1 \\ \phi_{11.0} + \phi_{00.1} - 1 \\ -\phi_{01.0} - \phi_{10.0} + \phi_{11.0} - \phi_{10.1} - \phi_{11.1} \\ -\phi_{10.0} - \phi_{11.0} - \phi_{01.1} - \phi_{10.1} + \phi_{11.1} \\ -\phi_{01.1} - \phi_{10.1} \\ -\phi_{01.0} - \phi_{10.0} \\ -\phi_{00.0} - \phi_{01.0} + \phi_{00.1} - \phi_{01.1} - \phi_{10.1} \\ \phi_{00.0} - \phi_{01.0} - \phi_{10.0} - \phi_{00.1} - \phi_{01.1} \end{array} \right\}, \quad (11.15)$$

$$\alpha^* := \min \left\{ \begin{array}{c} 1 - \phi_{10.0} - \phi_{01.1} \\ 1 - \phi_{01.0} - \phi_{10.1} \\ \phi_{00.0} - \phi_{01.0} + \phi_{11.0} + \phi_{00.1} + \phi_{01.1} \\ \phi_{00.0} + \phi_{01.0} + \phi_{00.1} - \phi_{01.1} + \phi_{11.1} \\ \phi_{00.1} + \phi_{11.1} \\ \phi_{00.0} + \phi_{11.0} \\ \phi_{10.0} + \phi_{11.0} + \phi_{00.1} - \phi_{10.1} + \phi_{11.1} \\ \phi_{00.0} - \phi_{10.0} + \phi_{11.0} + \phi_{10.1} + \phi_{11.1} \end{array} \right\}. \quad (11.16)$$

The above bounds are sharp in the sense that they can not be improved upon without making additional assumptions.

Example 11.5 Suppose we have observed $(\phi_{00.0}, \phi_{01.0}, \phi_{10.0}, \phi_{11.0}, \phi_{00.1}, \phi_{01.1}, \phi_{10.1}, \phi_{11.1}) = (0.919, 0, 0.081, 0, 0.315, 0.139, 0.073, 0.473)$. These values satisfy the instrumental inequalities (11.10). The bounds (11.15), (11.16) on α yield: $0.392 \leq \alpha \leq 0.780$. Indeed from (11.11)–(11.14) we obtain the stronger conclusion:

$$\begin{aligned} 0.081 &\leq \omega_0 \leq 0.081 \\ 0.473 &\leq \omega_1 \leq 0.861. \end{aligned}$$

In particular, in this special case we have achieved exact identification of ω_0 : $\Pr(Y = 1 \mid F_X = 0) = 0.081$. \square

In many applications there may be little or no data on all three variables (Z, X, Y) together, but (compare footnote 2 on page 81) we may have possibly separate sources of data on the pairs (Z, X) and (Z, Y) , thus allowing us to identify the observational distributions of $X \mid Z$ and $Y \mid Z$. These also can be used to supply bounds on causal effects (Ramsahai 2007). Using the values implied by those in Example 11.5, on the basis of this weaker information we obtain the weaker bounds $0.384 \leq \alpha \leq 0.919$.

11.4.3 General discrete variables

In principle the above methods can be extended to handle instrumental variable problems with discrete observed variables. In practice the inequalities become extremely complex and not particularly useful.

Chapter 12

Effect Of Treatment On The Treated

Formula (9.18) allows us to identify ACE from an observational study whenever we can also measure some sufficient covariate. But what can we do if we can not measure a sufficient covariate?

Definition 12.1 (Geneletti 2005). Let U be a sufficient covariate for the effect of T on Y . The *effect of treatment on the treated* (relative to U) is defined as:

$$\text{ETT}_U := \text{E}(\text{SCE}_U \mid T = 1, F_T = \emptyset). \quad (12.1)$$

□

Thus ETT_U is the average, in the observational regime, of the specific causal effect (relative to U) for those individuals who in fact receive the active treatment $T = 1$. This could be identified from observational data on all three variables (T, U, Y) . But now we assume that U is not observable, in any regime: a seemingly fatal handicap to identification of ETT_U .

12.1 Special cases

12.1.1 Allocation variable

An important special case of the above arises when, in the setting of §9.5, we take the sufficient covariate U to be the allocation indicator D (shown sufficient in §10.2.1). Then ETT_D is the average causal effect among those who would be treated if entered into the observational study.¹

12.1.2 Potential responses

As shown in §9.6, in a potential response setting we can formally treat the pair $U^* := (Y_0, Y_1)$ of potential responses as a sufficient covariate. Then (see §9.10.1) the associated specific causal effect is just the “individual causal effect”, $\text{ICE} \equiv Y_1 - Y_0$; and hence $\text{ETT}_{U^*} = \text{E}(Y_1 - Y_0 \mid T = 1, F_T = \emptyset)$. This potential response version is the usual definition of ETT. Expressed in this form it appears highly problematic, since in the observational regime we can never observe Y_0 when $T = 1$.

12.2 Uniqueness of ETT

Definition 12.1 appears, *prima facie*, to depend on the choice of the sufficient covariate U , and its distribution (jointly with the observables (T, Y)). However Theorem 12.1 below shows that this

¹A better term for ETT_D might be “the effect of treatment on the treatable”.

is not in fact the case: ETT_U does not depend on the choice of U , but only on the distributions of the observables Y and T in the different regimes.

Theorem 12.1 *For any sufficient covariate U ,*

$$\text{ETT}_U = \frac{\text{E}(Y \mid F_T = \emptyset) - \text{E}(Y \mid F_T = 0)}{\text{Pr}(T = 1 \mid F_T = \emptyset)}. \quad (12.2)$$

Proof. For $t = 0, 1$, define

$$k(t) := \text{E}\{\text{E}(Y \mid U, F_T = 0) \mid T = t, F_T = \emptyset\} \quad (12.3)$$

$$= \text{E}\{\text{E}(Y \mid U, T = 0, F_T = \emptyset) \mid T = t, F_T = \emptyset\} \quad (12.4)$$

by (9.8). In particular, $k(0) = \text{E}(Y \mid T = 0, F_T = \emptyset)$.

By (9.10) and conditional independence property (9.8), (12.1) becomes

$$\text{ETT}_U = \text{E}(Y \mid T = 1, F_T = \emptyset) - k(1). \quad (12.5)$$

Also,

$$\begin{aligned} \text{E}(Y \mid F_T = 0) &= \text{E}\{\text{E}(Y \mid U, F_T = 0) \mid F_T = 0\} \\ &= \text{E}\{\text{E}(Y \mid U, T = 0, F_T = \emptyset) \mid F_T = 0\} \\ &= \text{E}\{\text{E}(Y \mid U, T = 0, F_T = \emptyset) \mid F_T = \emptyset\} \end{aligned} \quad (12.6)$$

by (9.8) and (9.7). It follows that

$$\begin{aligned} \text{E}(Y \mid F_T = 0) &= \text{E}\{k(T) \mid F_T = \emptyset\} \\ &= \text{Pr}(T = 0 \mid F_T = \emptyset) k(0) + \text{Pr}(T = 1 \mid F_T = \emptyset) k(1). \end{aligned}$$

Hence

$$k(1) = \frac{\text{E}(Y \mid F_T = 0) - \text{Pr}(T = 0 \mid F_T = \emptyset) \text{E}(Y \mid T = 0, F_T = \emptyset)}{\text{Pr}(T = 1 \mid F_T = \emptyset)}. \quad (12.7)$$

Formula (12.2) now follows on substituting (12.7) into (12.5). \square

12.3 Identifying ETT

Theorem 12.1 reassures us that we could identify ETT if we could gather data on T under observational conditions, and on Y , both under observational conditions and for a random group of subjects assigned to the control treatment.

If we have only the observational data, we need to make further assumptions in order to identify ETT. One approach (Heckman and Robb 1985) involves modelling the way in which individuals make choices (*i.e.* the allocation indicator D): this is especially popular in Econometrics where it attempts to incorporate general principles of rational economic behaviour. Suitably strong model assumptions will allow observational identification of ETT.

Another approach is possible in the presence of an availability instrument Z , such that the situation can be modelled as in Figure 11.1 (with X now translated to T). Since $Z \perp\!\!\!\perp U$, the group assigned $Z = 0$, and thus forced to take $T = 0$, can be regarded as a random sample from the population. This group can thus serve as the interventional control group. The group assigned $Z = 1$ can similarly serve as the observational study group.

Chapter 13

Dynamic Treatment Strategies

The development in this chapter follows Dawid and Didelez (2005).

Suppose we can apply a sequence of treatments over time, observing interim “status” variables that we can then make use of in choosing the next treatment. Then we would like to assess the consequences of various *strategies* for making such sequential interventions on the basis of the sequentially accruing information. When and how can we do this from observational data?

Suppose that a patient presents with initial symptoms/status L_1 , and is then given a treatment T_1 . Following this he develops further symptoms/reactions L_2 , and is given a treatment T_2 , after which we observe the final response, Y . Any one of these variables may depend probabilistically on all the earlier ones. For notational simplicity we introduce $\bar{L}_i := (L_j : j \leq i)$, *etc.*

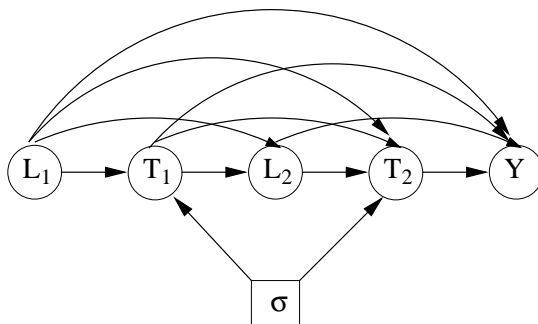


Figure 13.1: Sequential ignorability

Consider the influence diagram of Figure 13.1. Ignoring node σ , this is just a DAG model for the alternating sequence L_1, T_1, L_2, T_2, Y of status and treatment variables, allowing arbitrary probabilistic dependence.

The additional decision node σ indexes which regime is in operation. The possible values of σ range over $\mathcal{S}^* = \mathcal{S} \cup \{\emptyset\}$, where \mathcal{S} is the collection of interventional regimes we care about (*e.g.* we might want to optimize over $\sigma \in \mathcal{S}$) and \emptyset denotes the observational regime. The conditional distribution $p(t_i | \bar{l}_i, \bar{t}_{i-1}, \sigma)$ under any $\sigma \in \mathcal{S}$ (though not for $\sigma = \emptyset$) is supposed fully specified, describing the way (typically deterministic, but in principle possibly randomized) in which treatment T_i is to be chosen in response to the observed values of all earlier variables.

The significant property of Figure 13.1 is that there are no arrows out of σ into any status variable L_i (including $L_3 \equiv Y$). This is equivalent to the conditional independence properties

$$L_i \perp\!\!\!\perp \sigma \mid (\bar{L}_{i-1}, \bar{T}_{i-1}) \quad (i = 1, 2, 3). \quad (13.1)$$

That is to say, it is assumed that the conditional distribution of L_i , given the past and σ , does not depend on σ —in particular, it is the same for $\sigma = \emptyset$ as for any $\sigma \in \{\mathcal{S}\}$. In other words,

the probabilistic effect of all previous variables (both status and action) on a status variable is supposed to be a “stable feature”, the same across all regimes considered. This property is termed “sequential ignorability”. When it can be assumed, it permits (under a positivity condition) identification of $p(l_i | \bar{l}_{i-1}, \bar{t}_{i-1}, \sigma)$, for an interventional regime $\sigma \in \mathcal{S}$, from observational data collected under $\sigma = \emptyset$.

We are interested in discovering the distribution $p(y | \sigma)$ of the response Y that would result from applying some interventional regime σ . Simple probability theory gives:

$$\begin{aligned} p(l_1, t_1, l_2, t_2, y | \sigma) &= p(l_1 | \sigma) \\ &\quad \times p(t_1 | l_1, \sigma) \\ &\quad \times p(l_2 | l_1, t_1, \sigma) \\ &\quad \times p(t_2 | l_1, t_1, l_2, \sigma) \\ &\quad \times p(y | l_1, t_1, l_2, t_2, \sigma). \end{aligned}$$

Now because σ is a well-specified interventional regime, we know the (perhaps degenerate) distributions $p(t_1 | l_1, \sigma)$ and $p(t_2 | l_1, t_1, l_2, \sigma)$. Further, the sequential ignorability property $L_i \perp\!\!\!\perp \sigma | (\bar{L}_{i-1}, \bar{T}_{i-1})$ means that we can replace σ by \emptyset in the remaining terms $p(l_1 | \sigma)$, $p(l_2 | l_1, t_1, \sigma)$ and $p(y | l_1, t_1, l_2, t_2, \sigma)$. Hence we can evaluate the joint distribution of all the variables under σ . In particular, we can obtain the marginal distribution of the final response Y consequent on applying interventional strategy σ :

$$\begin{aligned} p(y | \sigma) &= \int dl_1 \int dt_1 \int dl_2 \int dt_2 \\ &\quad p(l_1 | \emptyset) \\ &\quad \times p(t_1 | l_1, \sigma) \\ &\quad \times p(l_2 | l_1, t_1, \emptyset) \\ &\quad \times p(t_2 | l_1, t_1, l_2, \sigma) \\ &\quad \times p(y | l_1, t_1, l_2, t_2, \emptyset). \end{aligned}$$

Moreover, we can do this for any $\sigma \in \mathcal{S}$.

A non-randomized interventional regime g is specified by functional relationships:

$$\left. \begin{aligned} T_1 &= g_1(L_1) \\ T_2 &= g_2(\bar{L}_2) \end{aligned} \right\} \quad (13.2)$$

corresponding to degenerate distributions. Then we obtain the *g-computation formula* (Robins 1986):

$$p(y | g) = \int \int p(l_1 | \emptyset) p(l_2 | L_1 = l_1, T_1 = g_1(l_1), \emptyset) p(y | L_1 = l_1, T_1 = g_1(l_1), L_2 = l_2, T_2 = g_2(\bar{l}_2), \emptyset) dl_1 dl_2. \quad (13.3)$$

The argument generalizes readily to $N > 2$ stages. The resulting formula is best expressed as *g-recursion*, a downwards recursion for $f(h) := p(y | h, \sigma)$, where h denotes some sequence of successive observations:

$$f(\bar{l}_i, \bar{t}_{i-1}) = \int p(t_i | \bar{l}_i, \bar{t}_{i-1}, \sigma) \times f(\bar{l}_i, \bar{t}_i) dt_i \quad (13.4)$$

$$f(\bar{l}_{i-1}, \bar{t}_{i-1}) = \int p(l_i | \bar{l}_{i-1}, \bar{t}_{i-1}, \emptyset) \times f(\bar{l}_i, \bar{t}_{i-1}) dl_i. \quad (13.5)$$

We start the recursion with

$$f(\bar{l}_N, \bar{t}_N) \equiv p(y | \bar{l}_N, \bar{t}_N, \sigma) = p(y | \bar{l}_N, \bar{t}_N, \emptyset),$$

and exit with the desired interventional distribution $p(y | \sigma)$.

For the special case that σ is a non-randomized strategy g , the dependence on the $t_i = g_i := g_i(\bar{l}_i)$ can be left implicit, and g -recursion becomes:

$$f(\bar{l}_{i-1}) = \int p(l_i | \bar{l}_{i-1}, \bar{g}_{i-1}, \emptyset) \times f(\bar{l}_i) dl_i. \tag{13.6}$$

13.1 More structure

Just as with the simple no-confounding property (9.3), we will often not find it easy to think directly about the acceptability of the sequential ignorability property (13.1). Rather, we might be able to build an acceptable conditional independence model to describe an extended situation, involving additional variables (possibly unmeasured). We can then use algebraic or graphical manipulation to investigate whether the desired conclusions $L_i \perp\!\!\!\perp \sigma | (\bar{L}_{i-1}, \bar{T}_{i-1})$ follow from the assumptions made.

For example, suppose that variables U_i , arising before the point of applying T_i , may possibly affect that and later treatments in the observational regime (though the U_i are *not* taken into account in the interventional strategies of interest). When can we ignore their presence, and still have sequential ignorability with respect to the (L_i) ? Two alternative sets of sufficient conditions are:

$$L_i \perp\!\!\!\perp \bar{U}_{i-1} | (\bar{L}_{i-1}, \bar{T}_{i-1}, \sigma), \tag{13.7}$$

a sequential version of (9.13); and

$$T_i \perp\!\!\!\perp \bar{U}_i | (\bar{L}_i, \bar{T}_{i-1}, \sigma), \tag{13.8}$$

a sequential version of (9.14).

For the case $N = 2$ these properties can be represented, respectively, by the IDs of Figure 13.2 and Figure 13.3, from which (13.1) can be verified by moralization.

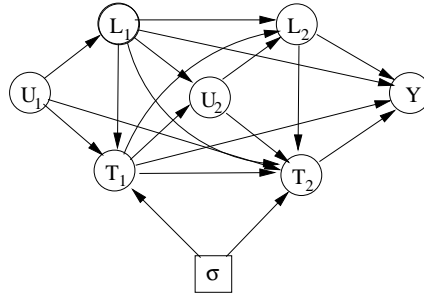


Figure 13.2: Sequential irrelevance

Under such conditions we can again apply g -computation to compute the effect of an interventional strategy that depends only on past (L_i) from observational data.

13.2 Other conditions

Conditions (13.7) and (13.8), while sufficient, are not necessary for sequential ignorability, and hence g -computation. Indeed, so long as we are only interested in computing the marginal interventional distribution of Y , rather than the joint interventional distribution of all variables, we even get by without sequential ignorability, while still being able to use g -computation (Pearl and Robins 1995; Dawid and Didelez 2005).

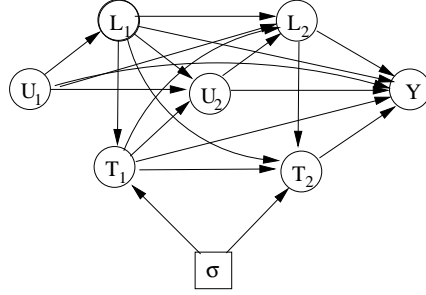


Figure 13.3: Sequential randomization

13.3 PR approach

For completeness we here describe the alternative conception of “sequential ignorability” within the potential response framework (Robins 1986; Robins 1987). This applies to non-randomized strategies g as in (13.2). Corresponding to any such strategy we are required to conceive of potential versions of all the status variables, (L_{1g}, L_{2g}, Y_g) . As is typical of the PR approach, all such potential variables, for all strategies g under consideration, are regarded as having simultaneous existence and a joint probability distribution.

We need to relate the potential variables to the actual observed variables (L_1, T_1, L_2, T_2, Y) , which is done as follows. Suppose that, in the observational regime, we are about to observe some status variable (*i.e.* L_1 , L_2 , or Y). Suppose further that, for any *earlier* treatment variable, its observed value happens to be the same as that prescribed by strategy g , applied to the information available at that decision point. Then we assume that that realized value of the upcoming status variable will be identical with its potential value under the operation of g . Mathematically, this is stated as:

$$\left. \begin{array}{l} T_1 = g(L_1) \\ T_1 = g(L_1), T_2 = g(L_1, L_2) \end{array} \right\} \begin{array}{l} L_1 = L_{1g} \\ \Rightarrow L_2 = L_{2g} \\ \Rightarrow Y = Y_g. \end{array} \quad (13.9)$$

Having set up this framework, the PR sequential ignorability condition, allowing identification of the “causal effect” of strategy g by means of the g -computation formula (13.3), is:

In the observational regime, each treatment variable is independent of all future potential observables under strategy g , given any history to date that is a possible history under g .

This is expressed mathematically as:

$$\left. \begin{array}{l} T_1 \perp\!\!\!\perp (L_{2g}, Y_g) \mid L_1 = l_1 \\ T_2 \perp\!\!\!\perp Y_g \mid L_1 = l_1, T_1 = g(l_1), L_2 = l_2 \end{array} \right\} \quad (13.10)$$

and can be regarded as a generalisation of the “no confounding” condition (9.6) to this dynamic setting. It can be shown (Dawid and Didelez 2005) that the above conditions imply those of (13.1).

Bibliography

- Balke, A. A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, (ed. R. L. de Mantaras and D. Poole), pp. 46–54.
- Balke, A. A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**, 1172–6.
- Bell, R. (1977). A reinterpretation of the direction of effects in studies of socialization. *Psychological Review*.
- Bell, R. Q. and Harper, L. V. (1977). *Child Effects on Adults*. Erlbaum, Hillsdale, N.J.
- Bjelakovic, G., Nikolova, D., Simonetti, R. G., and Gluud, C. (2004). Antioxidant supplements for preventing gastrointestinal cancers. *Cochrane Database of Systematic Reviews*. doi:10.1002/14651858.CD004183.pub2.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.
- Dawid, A. P. (1979a). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1–31.
- Dawid, A. P. (1979b). Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B*, **41**, 249–52.
- Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.
- Dawid, A. P. (1982). Intersubjective statistical models. In *Exchangeability in Probability and Statistics*, (ed. G. Koch and F. Spizzichino), pp. 217–32. North-Holland, Amsterdam.
- Dawid, A. P. (1985). Probability, symmetry and frequency. *British Journal for the Philosophy of Science*, **36**, 107–28.
- Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, **95**, 407–48.
- Dawid, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In *Highly Structured Stochastic Systems*, (ed. P. J. Green, N. L. Hjort, and S. Richardson), pp. 45–81. Oxford University Press.
- Dawid, A. P. and Didelez, V. (2005). Identifying the consequences of dynamic treatment strategies. Research Report 262, Department of Statistical Science, University College London.
- Dawid, A. P. and Evett, I. W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Science*, **42**, 226–31.
- de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, **7**, 1–68. Translated by H. E. Kyburg in Kyburg and Smokler (1964), 55–118.
- de Finetti, B. (1975). *Theory of Probability (Volumes 1 and 2)*. John Wiley and Sons, New York. (Italian original Einaudi, 1970).
- Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–53.

- Geiger, D. and Pearl, J. (1990). On the logic of causal models. In *Uncertainty in Artificial Intelligence 4*, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 3–14. North-Holland, Amsterdam.
- Geneletti, S. G. (2005). *Aspects of Causal Inference in a Non-Counterfactual Framework*. PhD thesis, Department of Statistical Science, University College London.
- Heckerman, D. and Shachter, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, **3**, 405–30.
- Heckman, J. and Robb, R. (1985). Alternative methods for estimating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, (ed. J. Heckman and B. Singer), pp. 156–245. New York: Cambridge University Press.
- Highways Agency (1997). *West London Speed Camera Demonstration Project: Analysis of Accident Data 36 Months Before and 36 Months After Implementation*. London Accident Analysis Unit, London.
- Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology*, **5**, (28).
doi:10.1186/1471-2288-5-28
<http://www.biomedcentral.com/1471-2288/5/28>.
- Holland, P. W. (1986). Statistics and causal inference (with Discussion). *Journal of the American Statistical Association*, **81**, 945–970.
- Howard, R. A. and Matheson, J. E. (1984). Influence diagrams. In *Readings in the Principles and Applications of Decision Analysis*, (ed. R. A. Howard and J. E. Matheson). Strategic Decisions Group, Menlo Park, California.
- Huang, Y. and Valtorta, M. (2006). Pearl’s calculus of intervention is complete. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, Arlington, Virginia.
- Kyburg, H. E. and Smokler, H. E. (ed.) (1964). *Studies in Subjective Probability*. John Wiley and Sons, New York.
- Lauritzen, S. L. (2000). Causal inference from graphical models. In *Complex Stochastic Systems*, (ed. O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg), chapter 2, pp. 63–107. CRC Press, London.
- Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.
- Lauritzen, S. L. and Nilsson, D. (1999). LIMIDs of decision problems. Technical Report R-99-2024, Department of Mathematical Sciences, Aalborg University.
- McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association*, **272**, 859–66.
- Nilsson, D. and Lauritzen, S. L. (2000). Evaluating influence diagrams using LIMIDs. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, (ed. C. Boutilier and M. Goldszmidt), pp. 436–45. Morgan Kaufmann Publishers, San Francisco.
- Oliver, R. M. and Smith, J. Q. (ed.) (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. John Wiley and Sons, Chichester, United Kingdom.
- Parascandola, M. and Weed, D. L. (2001). Causation in epidemiology. *Journal of Epidemiology and Community Health*, **55**, 905–12.
doi:10.1136/jech.55.12.905.
- Pearl, J. (1986). A constraint-propagation approach to probabilistic reasoning. In *Uncertainty in Artificial Intelligence*, (ed. L. N. Kanal and J. F. Lemmer), pp. 357–70. North-Holland, Amsterdam.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, San Mateo, California.

- Pearl, J. (1995a). Causal diagrams for empirical research (with Discussion). *Biometrika*, **82**, 669–710.
- Pearl, J. (1995b). Causal inference from indirect experiments. *Artificial Intelligence in Medicine*, **7**, 561–82.
- Pearl, J. (2000). *Causality*. Cambridge University Press.
- Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, (ed. P. Besnard and S. Hanks), pp. 444–53. Morgan Kaufmann Publishers, San Francisco.
- Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., Limacher, M., and Wactawski-Wende, J. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the Women’s Health Initiative clinical trial. *American Journal of Epidemiology*, **162**, (5), 404–14.
- Raiffa, H. (1968). *Decision Analysis*. Addison-Wesley, Reading, Massachusetts.
- Ramsahai, R. (2007). Causal bounds and instruments. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pp. 310–7. AUAI Press, Arlington, Virginia.
- Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, **30**, 962–1030.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–512.
- Robins, J. M. (1987). Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect”. *Comp. Math. Appl.*, **14**, 923–45.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–90.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The rôle of randomization. *Annals of Statistics*, **6**, 34–68.
- Rubin, D. B. (1980). Comment on “Randomization analysis of experimental data: The Fisher randomization test”, by D. Basu. *Journal of the American Statistical Association*, **81**, 961–2.
- Rubin, D. B. (1986). Which Ifs have causal answers? (Comment on “Statistics and causal inference”, by P. W. Holland). *Journal of the American Statistical Association*, **81**, 961–2.
- Rubin, D. B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, **26**, 20–36.
- Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research*, **34**, 871–82.
- Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 437–44. AUAI Press, Corvallis, Oregon.
- Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pp. 1219–26. AAAI Press, Menlo Park, California.
- Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search*. Springer-Verlag, New York.
- Stone, M. (2004). Radio 4 Today Programme Speed Tribunal. Research Report 245, Department of Statistical Science, University College London.

- Studený, M. (1989). Multi-information and the problem of characterization of conditional independence relations. *Problems of Control and Information Theory*, **18**, 3–16.
- Studený, M. (1992). Conditional independence relations have no finite complete characterization. In *Transactions of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 377–96. Academia, Prague.
- van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.
- Verma, T. and Pearl, J. (1990). Causal networks: Semantics and expressiveness. In *Uncertainty in Artificial Intelligence 4*, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 69–76. North-Holland, Amsterdam.
- Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence 6*, (ed. P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer), pp. 255–68. North-Holland, Amsterdam.
- Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, **2**, 121–34.