CAUSE AND CORRELATION IN BIOLOGY

A User's Guide to Path Analysis, Structural Equations and Causal Inference with R

Second Edition

Bill Shipley

Université de Sherbrooke, Canada

CAMBRIDGE
UNIVERSITY PRESS

1.1 The shadow's cause

The *Wayang Kulit* is an ancient theatrical art, practised in Malaysia and throughout much of the Orient. The stories are often about battles between good and evil, as told in the great Hindu epics. What the audience actually see are not actors, nor even puppets, but, instead, the shadows of puppets projected onto a canvas screen. Behind the screen is a light. The puppet master creates the action by manipulating the puppets and props so that they will intercept the light and cast shadows. As these shadows dance across the screen the audience must deduce the story from these two-dimensional projections of the hidden three-dimensional objects. However, shadows can be ambiguous. In order to imply the three-dimensional action, the shadows must be detailed, with sharp contours, and they must be placed in context.

Biologists are unwitting participants in nature's shadow play. These shadows are cast when the causal processes in nature are intercepted by our measurements. Like the audience at the *Wayang Kulit*, the biologist cannot simply peek behind the screen and directly observe the actual causal processes. All that can be directly observed are the consequences of these processes in the form of complicated patterns of association and independence in the data. As with shadows, these correlational patterns are incomplete – and potentially ambiguous – projections of the original causal processes. As with shadows, we can infer much about the underlying causal processes if we can learn to study their details and sharpen their contours, especially if we can study them in context.

Unfortunately, unlike the puppet master in a *Wayang Kulit*, who takes care to cast informative shadows, nature is indifferent to the correlational shadows that it casts. This is the main reason why researchers go to such extraordinary lengths to randomise treatment allocations and to control variables. These methods, when they can be properly done, simplify the correlational shadows to manageable patterns that can be more easily mapped onto the underlying causal processes.

It is uncomfortably true, though rarely admitted in statistics texts, that many important areas of science are stubbornly impervious to experimental designs based on the randomisation of treatments to experimental units. Historically, the response to this embarrassing problem has been either to ignore it or to banish the very notion of causality from the language and to claim that the shadows dancing on the screen are all that exists. Ignoring a problem doesn't make it go away, and defining a problem out of existence doesn't make it so. We need to know what we can safely infer about causes from their observational shadows, what we can't infer and the degree of ambiguity that remains.

I wrote this book to introduce biologists to some very recent, and intellectually elegant, methods that help in the difficult task of inferring causes from observational data. Some of these methods, such as structural equation modelling (SEM), are well known to researchers in other fields, though largely unknown to biologists. Other methods, such as those based on causal graphs, are unknown to almost everyone but a small community of researchers. These methods help both to test pre-specified causal hypotheses and to help discover potentially useful hypotheses concerning causal structures.

This book has three objectives. First, it was written to convince biologists that inferring causes without randomised experiments is possible. If you are a typical reader then you are already more than a little sceptical. For this reason I devote the first two chapters to explaining why these methods are justified. The second objective is to produce a user's guide, devoid of as much jargon as possible, that explains how to use and interpret these methods. In the service of this second objective I will explain, when appropriate, how to do this using the open source statistical program R.[1] The third objective is to exemplify these methods using biological examples, taken mostly from my own research and from that of my students. Since I am an organismal biologist whose research deals primarily with plant physiological ecology, most of the examples will be from this area, but the extensions to other fields of biology should be obvious.

I came to these ideas unwillingly. In fact, I find myself in the embarrassing position of having claimed publicly that inferring causes without randomisation and experimental control is probably impossible and, if possible, is not to be recommended (Shipley and Peters 1990). I expressed such an opinion in the context of determining how the different traits of an organism interact as a causal system. I will return to this theme repeatedly in this book, because it is so basic to biology,[2] and yet it is completely unamenable to the one method that most modern biologists and statisticians would accept as providing convincing evidence of a causal relationship: the randomised experiment. However, even as I advanced the arguments in 1990, I was dissatisfied with the consequences that such arguments entailed. I was also uncomfortably aware of the logical weakness of such arguments; the fact that I did not know of any provably correct way of inferring causation without the randomised experiment did not mean that such a method cannot exist. In my defence, and beyond the folly of youth, I could point out that I was saying nothing original; such an opinion was (and still is) the position of most statisticians and biologists. This view is summed up in the mantra that is learned by almost every student who has ever taken an elementary course in statistics: *correlation does not imply causation*.

In fact, with few exceptions,[3] correlation does imply causation. If we observe a systematic relationship between two variables, and we have ruled out the likelihood that this is simply due to a random coincidence, then *something* must be causing this relationship. When the audience at a Malay shadow theatre see a solid round shadow on the screen they know that some three-dimensional object has cast it, though they may not know if the object is a ball or a rice bowl in profile. A more accurate sound bite for introductory statistics would be that a simple correlation implies an *unresolved* causal structure, since we cannot know which is the cause and which is the effect, or if both are common effects of other unmeasured variables.

Although correlation implies an unresolved causal structure the reverse is not true: causation implies a completely resolved correlational structure. By this I mean that, once a causal structure has been proposed, the complete pattern of correlation and partial correlation is unambiguously fixed. This point is developed more precisely in Chapter 2, but it is so central to this book that it deserves repetition: the causal relationships between objects or variables determine the correlational relationships between them. Just as the shape of an object fixes the shape of its shadow, the patterns of direct and indirect causation fix the correlational ‘shadows’ that we see in observational data. The causal processes generating our observed data impose constraints on the patterns of correlation that such data display. This is the central insight underlying the methods described in this book.

The term ‘correlation’ evokes the notion of a probabilistic association between random variables. One reason why statisticians rarely speak of causation, except to distance themselves from it, is that there did not exist, until very recently, any rigorous translation between the language of causality (however defined) and the language of probability distributions (Pearl 1988). It is therefore necessary to link causation to probability distributions in a very precise way. Such rigorous links are now being forged. It is now possible to give mathematical proofs that specify the correlational pattern that must exist given a causal structure. These proofs also allow us to specify the class of causal structures that must include the causal structure that generates a given correlational pattern. The methods described in this book are justified by these proofs. Since my objective is to describe these methods and show how they can help biologists in practical applications, I won't present these proofs but will direct the interested reader to the relevant primary literature as each proof is needed.

Another reason why some prefer to speak of associations rather than causes is perhaps that causation is seen as a metaphysical notion that is best left to philosophers. In fact, even philosophers of science cannot agree on what constitutes a ‘cause’. I have no formal training in the philosophy of science and am neither able nor inclined to advance such a debate. This is not to say that philosophers of science have nothing useful to contribute. When it is directly relevant I will outline the development of philosophical investigations into the notion of ‘causality’ and place these ideas into the context of the methods that I will describe. However, I won't insist on any formal definition of ‘cause’, and will even admit that I have never seen anything in the life sciences that resembles the ‘necessary and sufficient’ conditions for causation that are so beloved of logicians.

You probably already have your own intuitive understanding of the term ‘cause’. I won't take it away from you, though I hope it will be more refined after reading this book. When I first came across the idea that one can study causes without defining them, I almost stopped reading the book (Spirtes, Glymour and Scheines 1993). I can advance three reasons why you should not follow through on this same impulse. First, and most important, the methods described here are not logically dependent on any particular definition of causality. The most basic assumption that these methods require is that causal relationships exist in relation to the phenomena that are studied by biologists.[4]

The second reason why you should continue reading even if you are sceptical is more practical and, admittedly, rhetorical: scientists commonly deal with notions whose meaning is somewhat ambiguous. Biologists are even more promiscuous than most with one notion that can still raise the blood pressure of philosophers and statisticians. This notion is ‘probability’, for which there are frequentist, objective Bayesian and subjective Bayesian definitions. In the 1920s von Mises is reported to have said: ‘Today, probability theory is not a mathematical science’ (Rao 1984). Mayo (1996) gives the following description of the present degree of consensus concerning the meaning of ‘probability’: ‘Not only was there the controversy raging between the Bayesians and the error [i.e. frequentist] statisticians, but philosophers of statistics of all stripes were full of criticisms of Neyman–Pearson error [i.e. frequentist-based] statistics.’ Needless to say, the fact that those best in a position to produce a definition of ‘probability’ cannot agree on one does not prevent biologists from effectively using probabilities, significance levels, confidence intervals and the other paraphernalia of modern statistics.[5] In fact, insisting on such an agreement would mean that modern statistics could not even have begun.

The third reason why you should continue reading, even if you are sceptical, is eminently practical. Although the randomised experiment is inferentially superior to the methods described in this book when randomisation can be properly applied, it cannot be properly applied to many (perhaps most) research questions asked by biologists. Unless you are willing simply to deny that causality is a meaningful concept then you will need some way of studying causal relationships when randomised experiments cannot be performed. Maintain your scepticism if you wish, but grant me the benefit of your doubt. A healthy scepticism while in a car dealership will keep you from buying a lemon. An unhealthy scepticism might prevent you from obtaining reliable transportation.

I said that the methods in this book are not logically dependent on any particular definition of causality. Rather than *defining* causality, the approach is to *axiomise* causality (Spirtes, Glymour and Scheines 1993). In other words, one begins by determining those attributes that scientists view as necessary for a relationship to be considered 'causal' and then develops a formal mathematical language that is based on such attributes. First, these relationships must be *transitive*: if A causes B and B causes C then it must also be true that A causes C. Second, such relationships must be 'local'; the technical term for this is that the relationships must obey the *Markov condition*, of which there are local and global versions. This is described in more detail in Chapter 2, but it can be intuitively understood to mean that events are caused only by their proximate causes. Thus, if event A causes event C *only* through its effect on an intermediate event B (A→B→C) then the causal influence of A on C is blocked if event B is prevented from responding to A. Third, these relationships must be *irreflexive*: an event cannot cause itself. This is not to say that every event must be causally explained; to argue in this way would lead us directly into the paradox of infinite regress. Every causal explanation in science includes events that are accepted (measured, observed…) without being derived from previous events.[6] Finally, these relationships must be *asymmetric*: if A is a cause of B, B cannot simultaneously be a cause of A.[7] In my experience, scientists generally accept these four properties. In fact, so long as I avoid asking for definitions, I find that there is a large degree of agreement between scientists on whether any particular relationship should be considered causal or not. It might be of some comfort to empirically trained biologists that the methods described in this book are based on an almost empirical approach to causality. This is because deductive definitions of philosophers are replaced with attributes that working scientists have historically judged to be necessary for a relationship to be causal. However, this change of emphasis is, by itself, of little use.

Next, we require a new mathematical language that is able to express and manipulate these causal relationships. This mathematical language is that of directed graphs[8] (Pearl 1988; Spirtes, Glymour and Scheines 1993). Even this new mathematical language is not enough to be of practical use. Since, in the end, we wish to infer causal relationships from correlational data, we need a logically rigorous way of translating between the causal relationships encoded in directed graphs and the correlational relationships encoded in probability theory. Each of these requirements can now be fulfilled.

1.2 Fisher's genius and the randomised experiment

Since this book deals with causal inference from observational data, we should first look more closely at how biologists infer causes from experimental data. What is it about these experimental methods that allows scientists to speak comfortably about causes? What is it about inferring causality from non-experimental data that makes them squirm in their chairs? I will distinguish between two basic types of experiments: the controlled experiment and the randomised experiment. Although the controlled experiment takes historical precedence, the randomised experiment takes precedence in the strength of its causal inferences.

Fisher[9] described the principles of the randomised experiment in his classic *Design of Experiments* (Fisher 1926). Since he developed many of his statistical methods in the context of agronomy, let us consider a typical randomised experiment designed to determine if the addition of a nitrogen-based fertiliser can cause an increase in the seed yield of a particular variety of wheat. A field is divided into 30 plots of soil and the seed is sown. The treatment variable consists of the fertiliser, which is applied at either 0 or 20 kg/hectare. For each plot we place a small piece of paper in a hat. One-half of the pieces of paper have a ‘0’ and the other half have a ‘20’ written on them. After thoroughly mixing the pieces of paper, we randomly draw one for each plot to determine the treatment level that each plot is to receive. After applying the appropriate level of fertiliser independently to each plot, we make no further manipulations until harvest day, at which time we weigh the seed that is harvested from each plot.

The seed weight per plot is normally distributed within each treatment group. Those plots receiving no fertiliser produce 55 g of seed with a standard error of six. Those plots receiving 20 kg/hectare of fertiliser produce 80 g of seed with a standard error of six. Excluding the possibility that a very rare random event has occurred (with a probability of approximately 5×10^{-8}), we have very good evidence that there is a positive *association* between the addition of the fertiliser and the increased yield of the wheat. Here we see the first advantage of randomisation. By randomising the treatment allocation, we generate a sampling distribution that allows us to calculate the probability of observing a given result by chance if, in reality, there is no effect from the treatment. This helps us to distinguish between chance associations and systematic ones. Since one error that a researcher can make is to confuse a real difference with a difference due to sampling fluctuations, the sampling distribution allows us to calculate the probability of committing such an error.[10] However, Fisher, and many other statisticians (Kempthorpe 1979; Kendall and Stuart 1983),[11] go further by claiming that the process of randomisation allows us to differentiate between associations due to causal effects of the treatment and associations due to some variable that is a common cause both of the treatment and response variables. What allows us to move so confidently from this conclusion about an *association* (a ‘co-relation’) between fertiliser addition and increased seed yield to the claim that the added fertiliser actually *causes* the increased yield?

Given that two variables (X and Y) are associated, there can be only three elementary, but not mutually exclusive, causal explanations; either X causes Y, Y causes X or there are some other causes that are common to both X and Y. Here, I am making no distinctions between ‘direct’ and ‘indirect’ causes; I argue in Chapter 2 that such terms have no meaning except relative to the other variables in the causal explanation. Remembering that transitivity is a property of causes, to say that X causes Y does not exclude the possibility that there are intervening variables ($X \rightarrow Z_1 \rightarrow Z_2 \rightarrow \ldots \rightarrow Y$) in the causal chain between them. We can confidently exclude the possibility that the seed produced by the wheat caused the amount of fertiliser that was added. First, we already know the only cause of the amount of fertiliser that was added to any given plot: the number on the piece of paper that was drawn from the hat. Second, the fertiliser was added before the wheat plants began to produce seed.[12] What allows us to exclude the possibility that the observed association between fertiliser addition and seed yield is due to some unrecognised common cause of both? This was Fisher's genius; the treatments were randomly assigned to the experimental units (i.e. the plots with their associated wheat plants). By definition, such a random process ensures that the order in which the pieces of paper are chosen (and therefore the order in which the plots receive the treatment) is causally independent of any attributes of the plot, its soil or the plant at the moment of randomisation.

Let's retrace the logical steps. We began by asserting that, if there was a causal relationship between fertiliser addition and seed yield, there would also be a systematic relationship between these two variables in our data: *causation implies correlation*. When we observe a systematic relationship that cannot reasonably be attributed to sampling fluctuations, we conclude that there was some causal mechanism responsible for this association. Correlation does not necessarily imply a causal relationship from the fertiliser addition to the seed yield, but it does imply *some* causal relationship that is responsible for this association. There are only three such elementary causal relationships, and the process of randomisation has excluded two of them. We are left with the overwhelming likelihood that the fertiliser addition caused the increased seed yield. We cannot categorically exclude the two alternative causal explanations, since it is always possible that we were incredibly unlucky. Perhaps the random allocations resulted, by chance, in those plots that received the 20 kg/hectare of fertiliser having soil with a higher moisture-holding capacity or some other attribute that actually caused the increased seed yield? In any empirical investigation, experimental or observational, all we can do is to advance an argument that is beyond reasonable doubt, not a logical certainty.

The key role played by the process of randomisation seems to be what ensures, up to a probability that can be calculated from the sampling distribution produced by the randomisation, that no uncontrolled common cause of both the treatment and the response variables could produce a spurious association. Fisher said as much himself when he stated that randomisation 'relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed'. Is this strictly true? Consider again the possibility that soil moisture content affects seed yield. By randomly assigning the fertiliser to plots, we ensure that, *on average*, the treatment and control plots have soil with the same moisture content, therefore removing any chance correlation between the treatment received by the plot and its soil moisture.[13] But the number of attributes of the experimental units (i.e. the plots with their attendant soil and plants) is limited only by our imagination. Let's say that there are 20 different attributes of the experimental units that could cause a difference in seed yield. What is the probability that at least one of these was sufficiently concentrated, by chance, in the treatment plots to produce a significant difference in seed yield even if the fertiliser had no causal effect? If this probability is not large enough for you then I can easily posit 50 or 100 different attributes that could cause a difference in seed yield. Since there are a large number of potential causes of seed yield, the likelihood that at least one of them was concentrated, by chance, in the treatment plots is not negligible even if we had used many more than the 30 plots.

Randomisation therefore serves two purposes in causal inference. First, it ensures that there is no causal effect coming from the experimental units to the treatment variable or from a common cause of both. Second, it helps to reduce the likelihood in the sample of a chance correlation between the treatment variable and some other cause of the treatment, but doesn't completely remove it. To cite Howson and Urbach (1989: 152, emphasis in original): 'Whatever the size of the sample, two treatment groups are *absolutely certain* to differ in some respect, indeed, in infinitely many respects, any of which might, unknown to us, be causally implicated in the trial outcome. So randomisation cannot possibly guarantee that the groups will be free from bias by unknown nuisance factors [i.e. variables correlated with the treatment]. And since one obviously doesn't know what those unknown factors are, one is in no position to calculate the probability of such a bias developing either.' This should not be interpreted as a severe weakness of the randomised experiment in any practical sense, but it does emphasise that even the randomised experiment does not provide any automatic assurance of causal inference, free of subjective assumptions.

Equally important is what is not required by the randomised experiment. The logic of experimentation up to Fisher's time was that of the controlled experiment, in which it was crucial that all other variables be experimentally fixed to constant values;[14] see, for example, Feiblman (1972: 149). Fisher (1970) explicitly rejects this as an inferior method, pointing out that it is logically impossible to know if 'all other variables' have been accounted for. This is not to say that Fisher does not advocate physically controlling for other causes in addition to randomisation. In fact, he explicitly recommends that the researcher do this whenever possible. For instance, in discussing the comparison of plant yields of different varieties, he advises that they be planted in soil 'that appears to be uniform'. In the context of pot experiments he recommends that the soil be thoroughly mixed before putting it in the pots, that the watering be equalised, that the pots receive the same amount of light, and so on. The strength of the randomised experiment lies in the fact that we do not have to physically control – or even be aware of – other causally relevant variables in order to reduce (but not logically exclude) the possibility that the observed association is due to some unmeasured common cause in our sample.

Yet strength is not the same as omnipotence. Some readers will have noticed that the logic of the randomised experiment has, hidden within it, a weakness not yet discussed that severely restricts its usefulness to biologists; a weakness that is not removed even with an infinite sample size. In order to work, one must be able to randomly assign values of the hypothesised 'cause' to the experimental units independently of any attributes of these units. This assignment must be direct and not mediated by other attributes of the experimental units. However, a large proportion of biological studies involve relationships between different attributes of such experimental units.

In the experiment described above, the experimental units are the plots of ground with their wheat plants. The attributes of these units include those of the soil, the surrounding environment and the plants. Imagine that the researcher wants to test the following causal scenario: the added fertiliser increases the amount of nitrogen absorbed by the plant. This increases the amount of nitrogen-based photosynthetic enzymes in the leaves and therefore the net photosynthetic rate. The increased carbon fixation due to photosynthesis causes the increased seed yield (Figure 1.1).

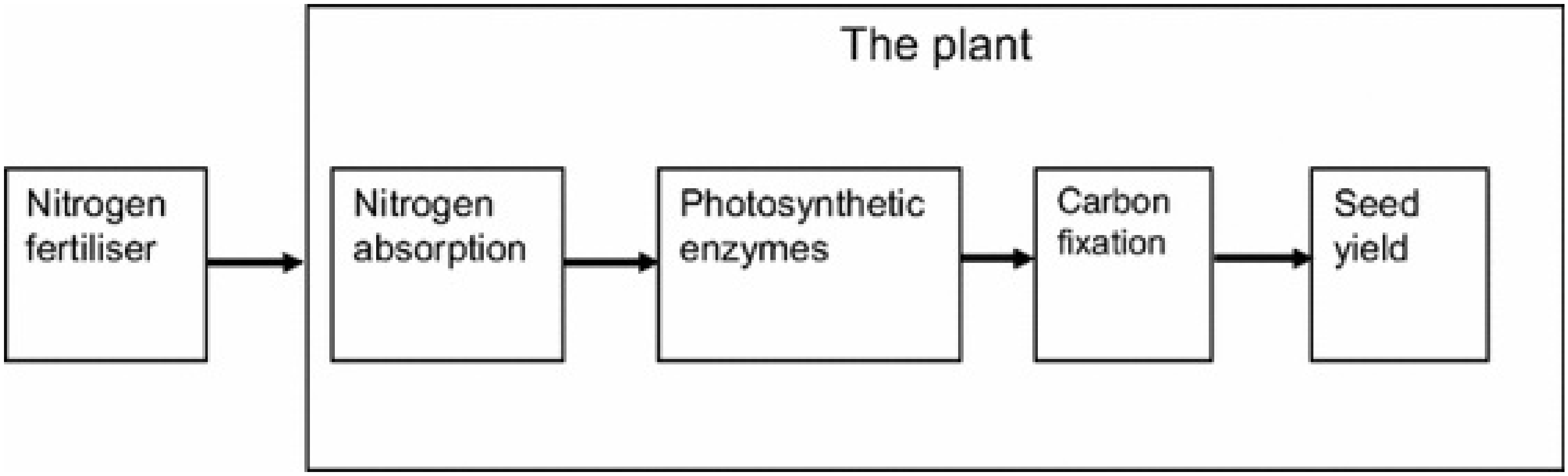


Figure 1.1 A hypothetical causal scenario that is not amenable to a randomised experiment

The first part of this scenario is perfectly amenable to the randomised experiment since the nitrogen absorption is an attribute of the plant (the experimental unit) while the amount of fertiliser added is controlled completely by the researcher independently of any attribute of the plot or its wheat plants. The rest of the hypothesis is impervious to the randomised experiment. For instance, both the rate of nitrogen absorption and the concentration of photosynthetic enzymes are attributes of the plant (the experimental unit). It is impossible to randomly assign rates of nitrogen absorption to each plant independently of any of its other attributes, yet this is the crucial step in the randomised experiment that allows us to distinguish correlation from causation. It is true that the researcher can induce a *change* both in the rate of nitrogen absorption by the plant and in the concentration of photosynthetic enzymes in its leaves, but in each case these changes are due to the addition of the fertiliser. After observing an association between the increased nitrogen absorption and the increased enzyme concentration the randomisation of fertiliser addition does not exclude different causal scenarios, only some of which are shown in Figure 1.2.

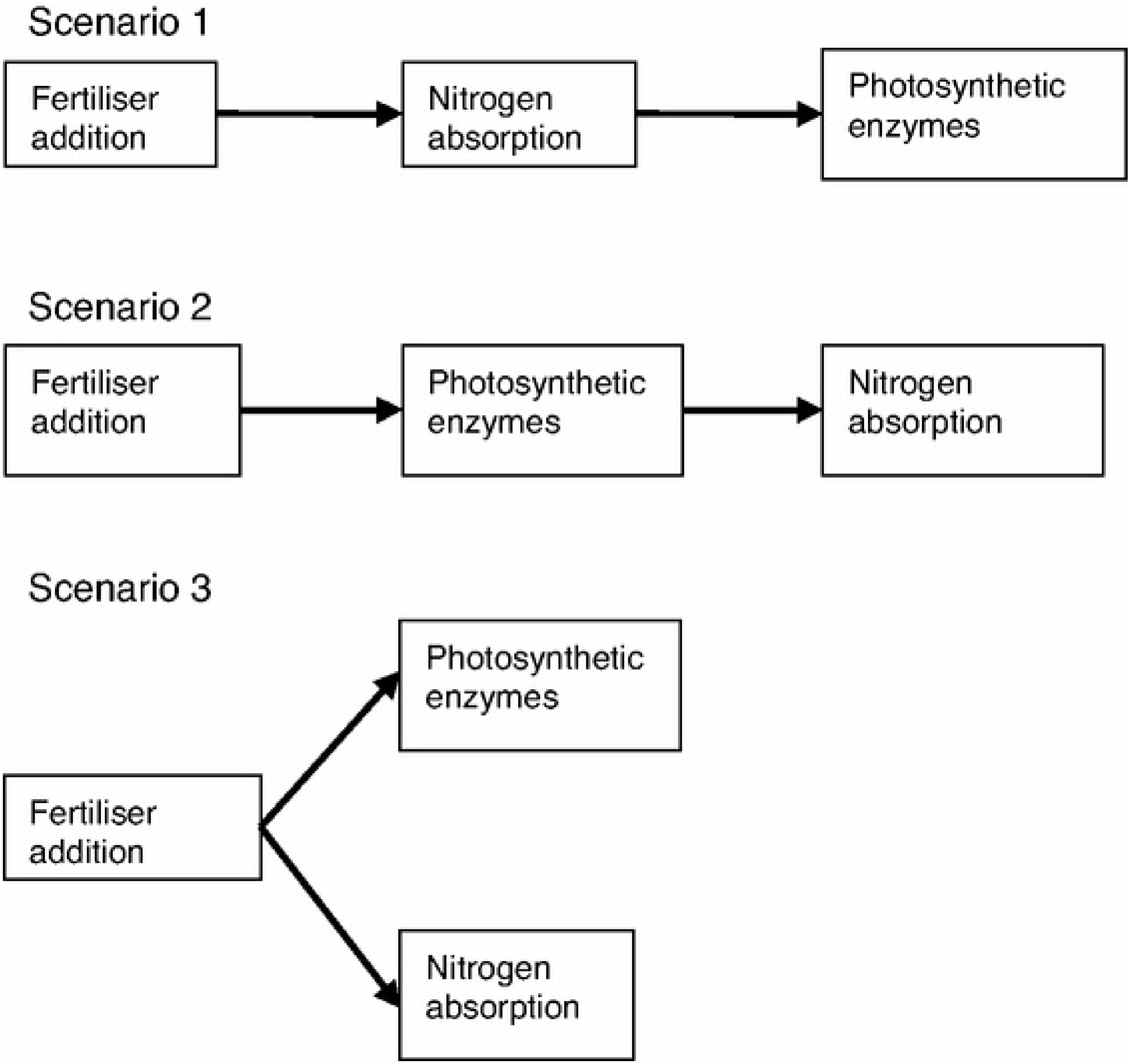


Figure 1.2 Three different causal scenarios that could generate an association between an increased nitrogen absorption and an increased enzyme concentration in the plant following the addition of fertiliser in a randomised experiment

When one is reading books about experimental design one's eyes often skim across the words 'experimental unit' without pausing to consider what these words mean. The experimental unit is the 'thing' to which the treatment levels are randomly assigned. The experimental unit is also an experimental *unit*. The causal relationships, if they exist, are between the external treatment variable and each of the attributes of the experimental unit that show a response. In biology, the experimental units (e.g. plants, leaves or cells) are integrated wholes whose parts cannot be disassembled without affecting the other parts. It is often not possible to randomly ‘assign’ values of one attribute of an experimental unit independently of the behaviour of its other attributes.[15] When such random assignments cannot be done, one cannot infer causality from a random experiment. A moment's reflection will show that this problem is very common in biology. Organismal, cell and molecular biology are rife with it. Physiology is hopelessly entangled. Evolution and ecology, dependent as they are on physiology and morphology, are often beyond its reach. If we accept that one cannot study causal relationships without the randomised experiment then a large proportion of biological research will have been gutted of any demonstrable causal content.

The usefulness of the randomised experiment is also severely reduced because of practical constraints. Remember that the inference is from the randomised treatment allocation to the experimental unit. The experimental unit must be the one that is relevant to the scientific hypothesis of interest. If the hypothesis refers to large-scale units (populations, ecosystems, landscapes) then the experimental unit must consist of such units. Someone wishing to know if increased carbon dioxide concentrations will change the community structure of forests will have to use entire forests as the experimental units. Such experiments are never done, and there is nothing in the inferential logic of randomised experiments that allows one to scale up from different (small-scale) experimental units. Even when proper randomised experiments can be done in principle, they cannot be done in practice due to financial or ethical constraints.

The biologist who wishes to study causal relationships using the randomised experiment is therefore severely limited in the questions that can be posed. The philosophically inclined scientist who insists that a positive response from a randomised experiment is an operational *definition* of a causal relationship would have to conclude that causality is irrelevant to much of science.

1.3 The controlled experiment

The currently prevalent notion that scientists cannot convincingly study causal relationships without the randomised experiment would have seemed incomprehensible to scientists before the twentieth century. Certainly, biologists *thought* that they were demonstrating causal relationships long before the invention of the randomised experiment. A wonderful example of this can be found in *An Introduction to the Study of Experimental Medicine* by the great nineteenth-century physiologist Claude Bernard (Bernard 1865).[16] I will cite a particularly interesting passage (Rapport and Wright 1963), and I ask that you pay special attention to the ways in which he tries to control variables. I will then develop the connection between the controlled experiment and the statistical methods described in this book.

> In investigating how the blood, leaving the kidney, eliminated substances that I had injected, I chanced to observe that the blood in the renal vein was crimson, while the blood in the neighbouring veins was dark like ordinary venous blood. This unexpected peculiarity struck me, and I thus made observation of a fresh fact begotten by the experiment, but foreign to the experimental aim pursued at the moment. I therefore gave up my unverified original idea, and directed my attention to the singular colouring of the venous renal blood; and when I had noted it well and assured myself that there was no source of error in my observation, I naturally asked myself what could be its cause. As I examined the urine flowing through the urethra and reflected about it, it occurred to me that the red colouring of the venous blood might well be connected with the secreting or active state of the kidney. On this hypothesis, if the renal secretion was stopped, the venous blood should become dark: that is what happened; when the renal secretion was re-established, the venous blood should become crimson again; this I also succeeded in verifying whenever I excited the secretion of urine. I thus secured experimental proof that there is a connection between the secretion of urine and the colouring of blood in the renal vein.

Our knowledge of human physiology has progressed far from the experiments of Bernard (physiologists might find it strange that he spoke of renal 'secretions'), yet his use of the controlled experiment would be immediately recognisable and accepted by modern physiologists. Fisher was correct in describing the controlled experiment as an inferior way of obtaining causal inferences, but the truth is that the randomised experiment is unsuited for much of biological research. The controlled experiment consists of proposing a hypothetical structure of cause–effect relationships, deducing what would happen if particular variables are controlled, or ‘fixed’ in a particular state, and then comparing the observed result with its predicted outcome. In the experiment described by Bernard, the hypothetical causal structure could be conceptualised as shown in Figure 1.3.

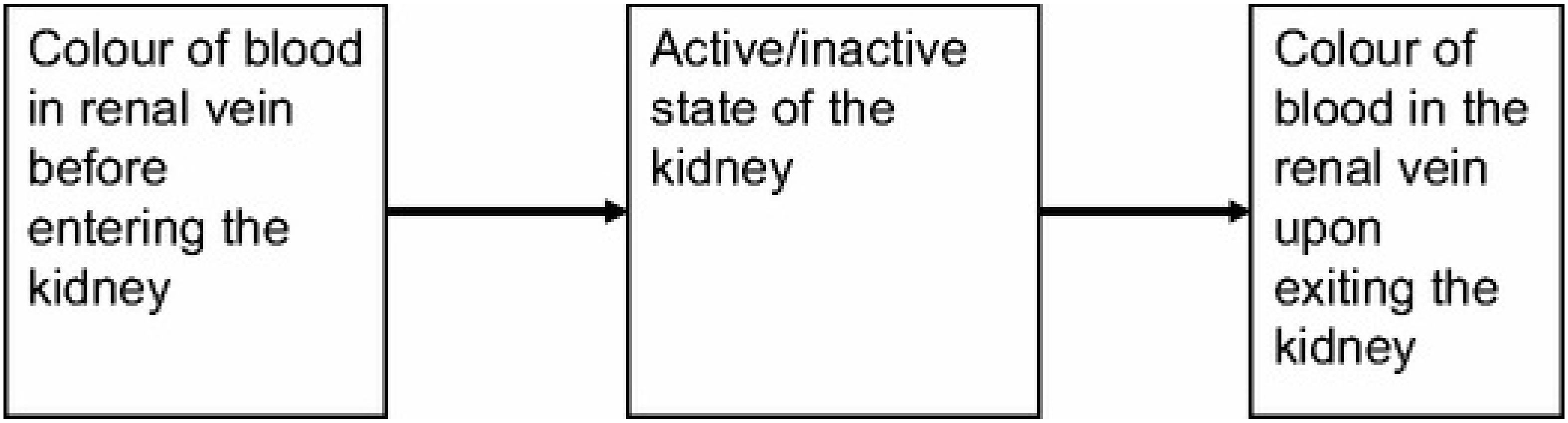

Colour of blood in renal vein before entering the kidney
Active/inactive state of the kidney
Colour of blood in the renal vein upon exiting the kidney

Figure 1.3 The hypothetical causal explanation invoked by Claude Bernard

The key notion in Bernard's experiment was the realisation that, if his causal explanation was true, the type of *association* between the colour of the blood in the renal vein as it enters and leaves the kidney would change depending on the state of the hypothesised cause – i.e. whether the kidney was secreting or not. It is worth returning to his words: ‘On this hypothesis, if the renal secretion was stopped, the venous blood should become dark: that is what happened; when the renal secretion was re-established, the venous blood should become crimson again; this I also succeeded in verifying whenever I excited the secretion of urine. I thus secured experimental proof that there is a connection between the secretion of urine and the colouring of blood in the renal vein.’ Since he had explicitly stated earlier in the quote that he was enquiring into the ‘cause’ of the phenomenon, it is clear that he viewed the result of his experiments as establishing a *causal connection* between the secretion of urine and the colouring of blood in the renal vein.

Although the controlled experiment is an inferior method of making causal inferences relative to the randomised experiment, it is actually responsible for most of the causal knowledge that science has produced. The method involves two basic parts. First, one must propose an hypothesis stating how the measured variables are linked in the causal process. Second, one must deduce how the associations between the observations must change once particular combinations of variables are controlled so that they can no longer vary naturally – i.e. once particular combinations of variables are ‘blocked’. The final step is to compare the patterns of association after such controls are established with the deductions. Historically, variables have been blocked by physically manipulating them. However (this is an important point that will be more fully developed and justified in Chapter

2), it is the control of variables, not how they are controlled, that is the crucial step. The weakness of the method, as Fisher pointed out, is that one can never be sure that all relevant variables have been identified and properly controlled. One can never be sure that, in manipulating one variable, one has not also changed some other, unknown variable. In any field of study, as Bernard documents in his book, the first causal hypotheses are generally wrong, and the process of testing, rejecting and revising them is what leads to progress in the field.

1.4 Physical controls and observational controls

It is the control of variables, not how they are controlled, that is the crucial step in the controlled experiment. What does it mean to 'control' a variable? Can such control be obtained in more than one way? In particular, can one control variables based on observational, rather than experimental, observations? The link between a physical control through an experimental manipulation and a statistical control through conditioning will be developed in the next chapter, but it is useful to provide an informal demonstration here using an example that should present no metaphysical problems to most biologists.

Body size in large mammals seems to be important in determining much of their ecology. In populations of bighorn sheep in the Rocky Mountains, it has been observed that the probability of survival of an individual through the winter is related to the size of the animal in the fall. However, this species has a strong sexual dimorphism, with males being up to 60 per cent larger than females. Perhaps the association between body size and survival is simply due to the fact that males have a better probability of survival than females, and this is unrelated to their body size? In observing these populations over many years, perhaps the observed association arises because those years showing better survival also have a larger proportion of males? Figure 1.4 shows these two alternative causal hypotheses. I have included boxes labelled 'Other causes' to emphasise that we are not assuming the chosen variables to be the only causes of body size or of survival.

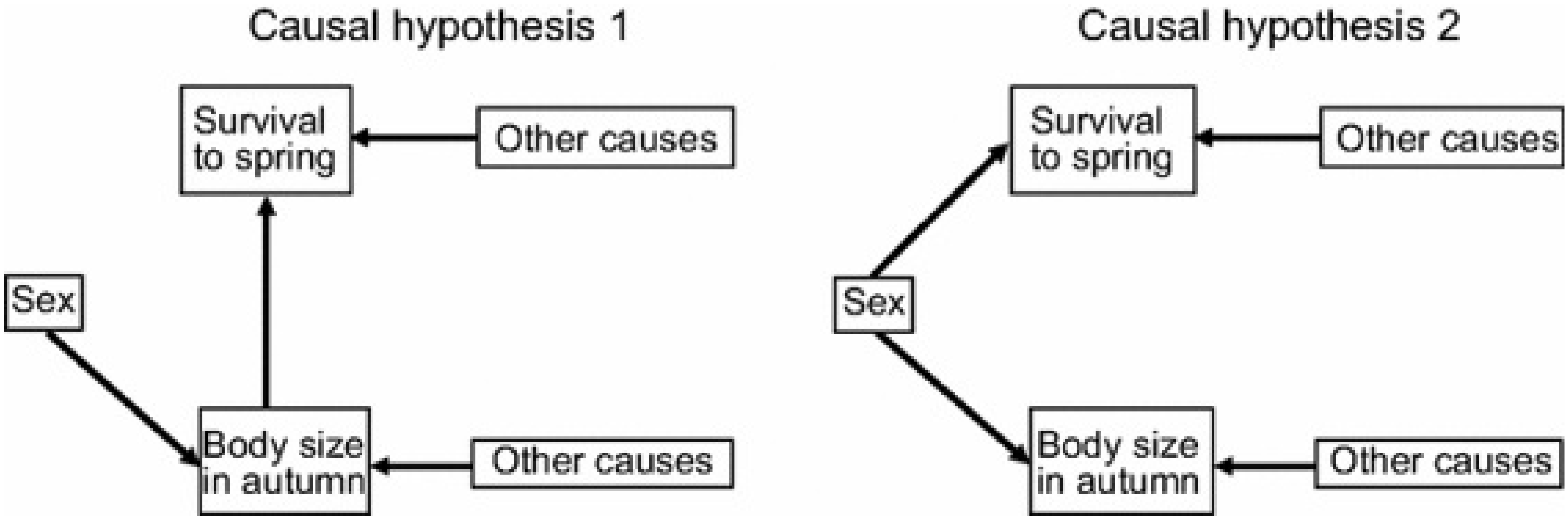


Figure 1.4 Two alternative causal explanations for the relationship between the sex and body size of bighorn sheep in the autumn and the probability of survival until the spring

Notice the similarity to Claude Bernard's question concerning the cause of blood colour in the renal vein. The difference between the two alternative causal explanations in Figure 1.4 is that the second assumes that the association between spring survival and autumn body size is due only to the sex ratio of the population. Thus, if the sex ratio could be held constant, the association would disappear. Since adult males and females of this species live in separate groups, it would be possible to physically separate them in their range and, in this way, physically control the sex ratio of the population. However, it is much easier to simply sort the data according to sex and then look for an association within each homogeneous group. The act of separating the data into two groups such that the variable in question – the sex ratio – is constant within each group represents a *statistical control*. We could imagine a situation in which we instruct one set of researchers to physically separate the original population into two groups based on sex, after which they test for the association within each of their experimental groups, and then ask them to combine the data together and give them to a second team of researchers. The second team would analyse the data using the statistical control. Both groups would come to identical conclusions.[17] In fact, using statistical controls might even be preferable in this situation. Simply observing the population over many years and then statistically controlling for the sex ratio on paper does not introduce any physical changes in the field population. It is likely that the act of physically separating the sexes in the field might introduce some unwanted, and potentially uncontrolled, change in the behavioural ecology of the animals during the rut that might bias the survival rates during the winter quite independently of body size.

Let's further extend this example to look at a case in which it is not as easy to separate the data into groups that are homogeneous with respect to the control variable. Perhaps the researchers have also noticed an association between the amount and quality of the rangeland vegetation during the early summer and the probability of survival during the next winter. They hypothesise that this pattern is caused by the animals being able to eat more during the summer, which increases their body size in the autumn, which in turn increases their chances of survival during the winter (Figure 1.5).

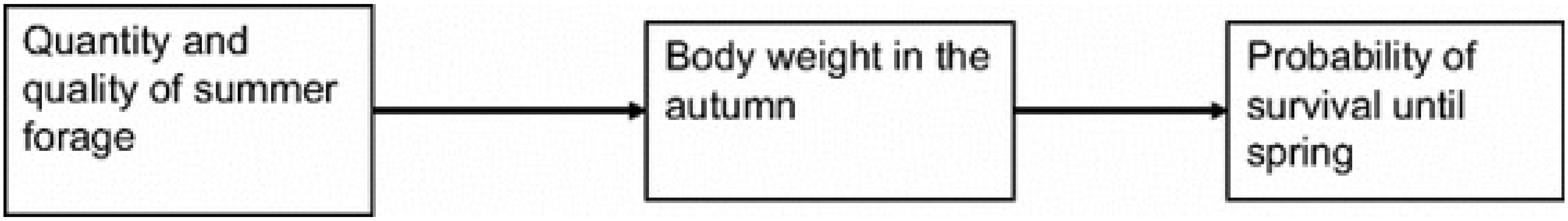

Quantity and quality of summer forage
Body weight in the autumn
Probability of survival until spring

Figure 1.5 A hypothetical causal explanation for the relationship between the quality and quantity of summer forage, the body weight of bighorn sheep in the autumn and the probability of survival until the spring

The logic of the controlled experiment requires us to be able to compare the relationship between forage quality and winter survival after physically preventing body weight from changing, which we cannot do.[18] Since body weight is a continuous variable, we can't simply sort the data and then divide it into groups that are homogeneous for this variable. This is because each animal will have a different body weight. Nonetheless, there is a way of comparing the relationship between forage quality and winter survival while controlling the body weight of the animals during the comparison. This involves the concept of statistical conditioning, which will be more rigorously developed in Chapters 2 and 3. An intuitive understanding can be had with reference to a simple linear regression (Figure 1.6).

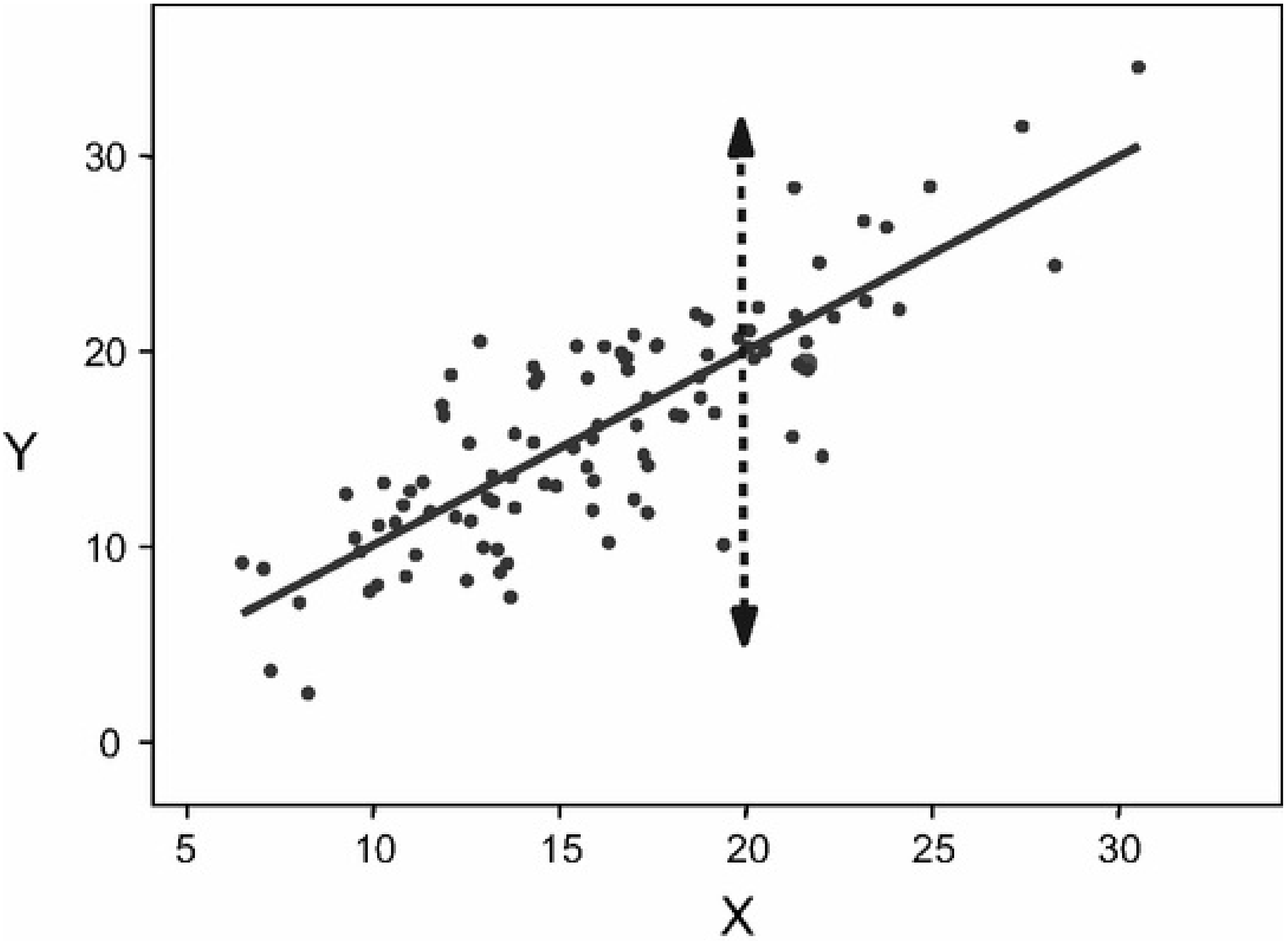


Figure 1.6 A simple bivariate regression

Notes: The solid line shows the expected value of y_i given the value of x_i ($E[y_i|x_i]$). The dotted line shows the possible values of y_i that are independent of x_i (the residuals).

The formula for a linear regression is $y_i = \alpha + \beta x_i + N(0, \sigma)$. Here, the notation $N(0, \sigma)$ means ‘a normally distributed random variable with a population mean of zero and a population standard deviation of σ’. As the formula makes clear, the observed value of y consists of two parts: one part that depends on x and one part that doesn't. If we let E(y|x) represent the expected value of y given x, we can write

$$
\begin{aligned}
&E(y|x_i) = \alpha + \beta x_i \\
&y_i = E(y|x_i) + N(0, \sigma) \\
&(y_i - E(y|x_i)) = N(0, \sigma)
\end{aligned}
$$

Thus, if we subtract the expected value of each y, given x, from the value itself, we get the variation in y that is independent of x. This new variable is called the *residual* of y given x. These are the values of y that exist for a constant value of x. For instance, the vertical arrow in Figure 1.6 shows the values of y when x = 20.

If we want to compare the relationship between forage quality and winter survival while controlling the body weight of the animals during the comparison then we have to remove the effect of body weight on each of the other two variables. We do this by taking each variable in turn, subtracting the expected value of it given body weight and then seeing if there is still a relationship between the two sets of residuals. In this way, we can hold constant the effect of body weight in a way similar to experimentally holding constant the effect of some variable. The analogy is not exact. There are situations in which statistically holding constant a variable will produce different patterns of association from those that would occur when physically holding constant the same variable. To understand when statistical controls cast the same correlational shadows as experimental controls, and when they differ, we need a way of rigorously translating from the language of causality to the language of probability distributions. This is the topic of the next chapter.

[1] See www.r-project.org.

[2] This is also the problem that inspired Sewall Wright, one the most influential evolutionary biologists of the twentieth century, the inventor of path analysis and the intellectual grandparent of the methods described in this book. The history of path analysis is explored in more detail in Chapter 3.

[3] It could be argued that variables that covary only because they are time-ordered have no causal basis.

[4] Perhaps quantum physics does not need such an assumption. I will leave this question to people better qualified than I. The world of biology does not operate at the quantum level.

[5] The perceptive reader will note that I have now compounded my problems. Not only do I propose to deal with one imperfectly defined notion – causality – but I will do it with reference to another imperfectly defined notion: a probability distribution.

[6] The paradox of infinite regress is sometimes 'solved' by simply declaring a first cause: that which causes but which has no cause. This trick is hardly convincing, because, if we are allowed to invent such things by fiat, then we can declare them anywhere in the causal chain. The antiquity of this paradox can been seen in the first sentence of the first verse of Genesis: 'In the beginning God created the heavens and the earth.' According to the Confraternity Text of the Holy Bible, the Hebrew word that has been translated as 'created' was used only with reference to divine creation and meant 'to create out of nothing'.

[7] This does not exclude feedback loops so long as we understand these to be dynamic in nature: A causes B at time t, B causes A at time t+Δt, and so on. This is discussed more fully in Chapter 2.

[8] Biologists will find it ironic that this graphical language was actually proposed by Wright (1921), one of the most influential evolutionary biologists of the twentieth century, but his insight was largely ignored. This history is explored in Chapters 3 and 4.

[9] Sir Ronald A. Fisher (1890–1962) was chief statistician at the Rothamsted Agricultural Station. He was later Galton Professor at the University of London and Professor of Genetics at the University of Cambridge.

[10] It is for this reason that Mayo (1996) calls such frequency-based statistical tests 'error probes'.

[11] 'Only when the treatments in the experiment are applied by the experimenter using the full randomisation procedure is the chain of inductive inference sound; it is only under these circumstances that the experimenter can attribute whatever effect he observes to the treatment and to the treatment only' (Kempthorpe 1979).

[12] Unless your meaning of 'cause' is very peculiar, you will not have objected to the notion that causal relationships cannot travel backwards in time. Despite some ambiguity in its formal definition, scientists would agree on a number of attributes associated with causal relationships. As with pornography, we have difficulty defining it but we all seem to know it when we see it.

[13] More specifically, these two variables, being causally independent, are also probabilistically independent in the statistical population. This is not necessarily true in the sample due to sampling fluctuations.

[14] Clearly, this cannot be literally true. Consider a case in which the causal process is A→B→C, and we want to experimentally test whether A causes C. If we hold variable B constant then we would incorrectly surmise that A has no causal effect on C. It is crucial that common causes of A and C be held constant in order to exclude the possibility of a spurious relationship. It is also a good idea, though not crucial for the causal inference, that causes of C that are independent of A also be held constant, in order to reduce the residual variation of C.

[15] This is not to say that it is always impossible. For instance, one can randomly add levels of insulin to the blood because the only cause of these changes (given proper controls) is the random numbers assigned to the animal. One cannot randomly add different numbers of functioning chloroplasts to a leaf.

[16] Rapport and Wright (1963) describe Claude Bernard (1813–1878) as an experimental genius and ‘a master of the controlled experiment’.

[17] It is not true that statistical and physical controls will always give the same conclusion. This is discussed in Chapter 2.

[18] It is actually possible, in principle if not in practice, to conduct a randomised experiment in this case, so long as we are interested only in knowing if summer forage quality causes a change in winter survival. This is because the hypothetical cause (vegetation quality and quantity) is not an attribute of the unit possessing the hypothetical effect (winter survival). Again, it is impossible to use a randomised experiment to determine if body size in the autumn is a cause of increased survival during the winter.

2

From cause to correlation and back

2.1 Translating from causal to statistical models

The official language of statistics is the probability calculus, based on the notion of a probability distribution. For instance, if you conduct an analysis of variance (ANOVA) then the key piece of information is the probability of observing a particular value of Fisher's F statistic in a random sample of data, given a particular hypothesis or model. To obtain this crucial piece of information, you (or your computer) must know the probability density function of the F statistic. Certain other (mathematical) languages are tolerated within statistics but, in the end, one must link one's ideas to a probability distribution in order to be understood. If we wish to study causal relationships using statistics, it is necessary that we translate, without error, from the language of causality to the only language that statistics can understand: probability theory.

Such a rigorous translation device did not exist until very recently (Pearl 1988). It is no wonder that statisticians have virtually banished the word ‘cause’ from statistics; such a word has no equivalent in their language.[1] Within the world of statistics the scientific notion of causality has, until recently, been a stranger in a strange land. Posing causal questions in the language of the probability calculus is like a unilingual Englishman asking for directions to the Louvre from a Frenchman who can't speak English. The Frenchman might understand that directions are being requested, and the Englishman might see fingers pointing in particular directions, but it is not at all certain that works of art will be found. Imperfect translations between the language of causality and the language of probability theory are equally disorienting.

Mistakes in translation come in all kinds. The most dangerous ones are the subtle errors in which a slight change in inflection or context of a word can change the meaning in disastrous ways. Because the French word *demande* both sounds like the English word ‘demand’ and has roughly the same meaning (it simply means ‘to ask for’, without any connotation of obligation), I have seen French-speaking people come up to a shop assistant and, while speaking English, ‘demand service’. They think that they are politely asking for help, while the assistant thinks they are issuing an ultimatum. I once came close to being beaten by an enraged boyfriend simply because (I thought) I was complimenting his girlfriend on her long hair, which was drawn in a ponytail. The word for ‘tail’ in French is *queue*, which takes a feminine gender. There is another word in colloquial French, *cul* (the ‘l’ is silent), that sounds almost the same. It takes a masculine gender, is pronounced only slightly differently and refers to a person's rear end; the correctly translated word rhymes with ‘pass’, but the reader will understand if I don't give the literal translation. So, while trying to make conversation with the boyfriend, I told him that his girlfriend had a nice *cul* instead of a nice *queue*. I immediately knew, from the look of rage on his face, that I had chosen the wrong word.

The same subtle mistakes of translation can occur when translating between the language of causality and the mathematical language of probability distributions. I began the first chapter by comparing causes and correlations to three-dimensional objects and their two-dimensional shadows. Clearly, there is a close relationship between the object and its shadow. Just as clearly, they are not the same thing. The goal of this chapter is to describe the relationship between variables involved in a causal process and the probability distribution of these variables that the causal process generates. Causal processes cast probability shadows, but 'causes' and 'probability distributions' are not the same thing either. It is important to understand exactly how the translation is made between causal processes and probability distributions in order to avoid the scientific equivalent of a punch in the nose from an enraged boyfriend.

I will make the distinction between a causal model, an observational model and a statistical model. Since every child knows that rain causes mud,[2] I will illustrate the difference between these three types of models with this analogy. The statement 'Rain causes mud' implies an asymmetric relationship: the presence of rain will create mud, but the presence of mud will not create rain. I will use the symbol '→' when I want to refer to such causal relationships. This leads naturally to the sort of 'box and arrow' diagrams with which most biologists are familiar (Figure 2.1).

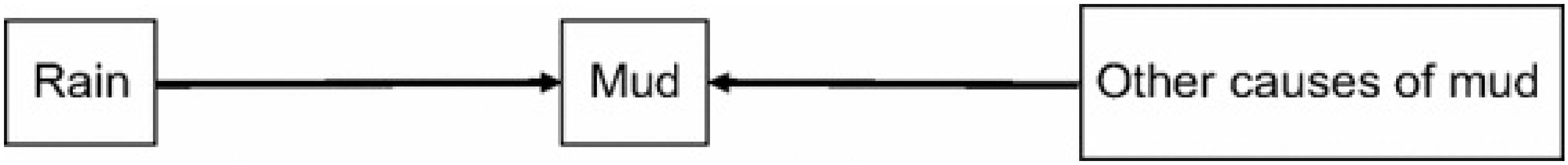


Figure 2.1 The causal relationships between rain, mud and other causes of mud

To complete the description it is necessary to add the convention that, unless a causal relationship is explicitly included, it is understood not to exist. So, in Figure 2.1, the fact that there are no arrows between 'Rain' and 'Other causes of mud' means that there is no direct causal relationship between them; in fact, there is no causal relationship of any kind in this example, since the two are causally independent.

The observational model that is related to this causal model is the statement 'Having observed rain will give us information about what we will observe concerning mud'. Notice that this observational statement deals with information, not causes, and is not asymmetric. If we learn that it has rained then we will have added information concerning the presence of mud in our yard, but observing mud in our yard will also give us information about whether or not it has rained. I will use the symbol ‘—’ when I refer to such observational relationships. This leads to the model in Figure 2.2.

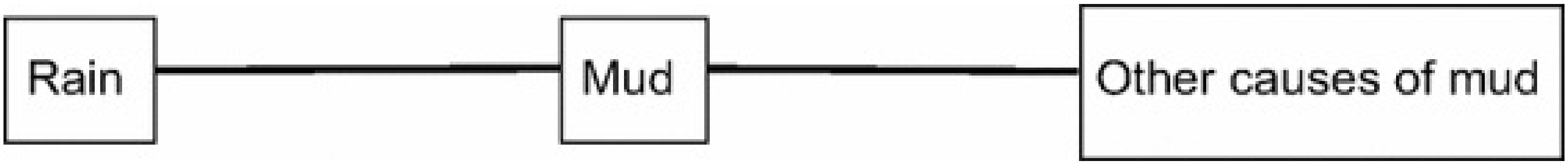


Figure 2.2 The observational relationships between rain, mud and other causes of mud

Notice that, although rain and other causes of mud are causally independent, they are not observationally independent given the state of mud; knowing that it has not rained but that there is mud in the front yard gives you information on the existence of other causes of mud.

The statistical model differs only in degree, not in kind, from the observational model. The statistical model (Figure 2.3) specifies the mathematical relationship between the variables as well as the probability distributions of the variables. Now we can use the equivalence operator of algebra (=), since we are stating a quantitative equivalence.

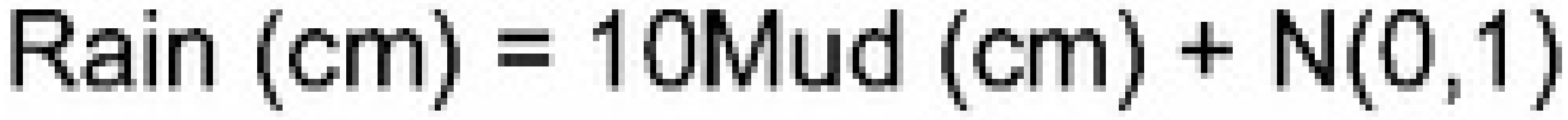
$$\text{Mud (cm)} = 0.1\text{Rain (cm)} + N(0{,}0.1)$$

Figure 2.3 A statistical model relating rain and mud

This mathematical statement says that the value obtained by measuring the depth of the mud, in centimetres, is the same as (is ‘equivalent to’) the value that is obtained by measuring the amount of rain that falls, in centimetres, multiplying this value by 0.1, and adding another value (in centimetres) obtained from a random value taken from a normal distribution whose population mean is zero and whose population standard deviation in 0.1.

What is the point of all this? According to Pearl (1997), a century of confusion between correlation and causation can be traced, in part, to a mistranslation of the word *cause.* When scientists and statisticians attempt to express notions of causality using mathematics they mistranslate ‘cause’, a word having connotations of asymmetry and all the other properties discussed in Chapter 1, as the algebraic notion “=” used in the language of probability theory. The symbols → and = do not mean the same thing, because the algebraic concept of ‘equivalence’ and its symbol (=) do not have the properties of causes discussed in Chapter 1. It is perfectly correct to rearrange the equation in

Figure 2.3 in order to imply that the amount of rain can be predicted from the amount of mud (Figure 2.4), even though any five-year-old child would recognise this as causally nonsensical.



$$\text{Rain (cm)} = 10\text{Mud (cm)} + N(0,1)$$

Figure 2.4 Another statistical model relating rain and mud

This mistake is the scientific equivalent of telling a boyfriend that his girlfriend has *un beau cul* rather than *une belle queue*. The conceptual error occurs because we have replaced → with =. After translating from the language of causality to the language of observations, we have used the syntax of this observational language to produce a perfectly reasonable statement for this observational language, but then we have performed a literal translation back into the language of causality without recognising the difference in syntax. There are computer programs that attempt to translate between human languages, and those that use literal word-by-word translations run into the same problems.[3] A newspaper headline such as ‘Bill Gates worth $1,000,000,000’, after being literally translated (word for word) into a different language and then retranslated back into English, might come up with a phrase such as ‘payment request [“bill”] for doors in the fence [“gates”] costs $1,000,000,000’!

In the next few sections I develop a translation device to move between causal models and observational (statistical) models. To do this we require the necessary and sufficient conditions needed to specify a joint probability distribution that must exist given a causal process. Put another way, we require the necessary and sufficient conditions needed to specify the correlational shadow that will be cast by a causal process. This provides the key to translating between causal and statistical models. These sections require more effort to understand but in each case I will also provide a more intuitive description and some worked examples.

The strategy for translation from the physical world, in which the notion of causation is applicable, to the mathematical world of probability theory, in which the abstract notion of algebraic equivalence is applicable, involves two steps. First, since algebra cannot express the sort of relationships that we consider ‘causal’, we need a new mathematical language that can; this language is that of directed graphs. Second, we need a translation device that can unambiguously convert the statements expressed in such directed graphs into statements concerning the conditional independence of random variables obeying a particular probability distribution. This translation device is called ‘d-separation’ (short for ‘*directed* separation’).

2.2 Directed graphs

It is now time to introduce some terminology concerning directed (sometimes called causal) graphs. These terms, although unfamiliar to most biologists, are quite easy to grasp and use. These terms will be defined using the causal graph shown in Figure 2.5.

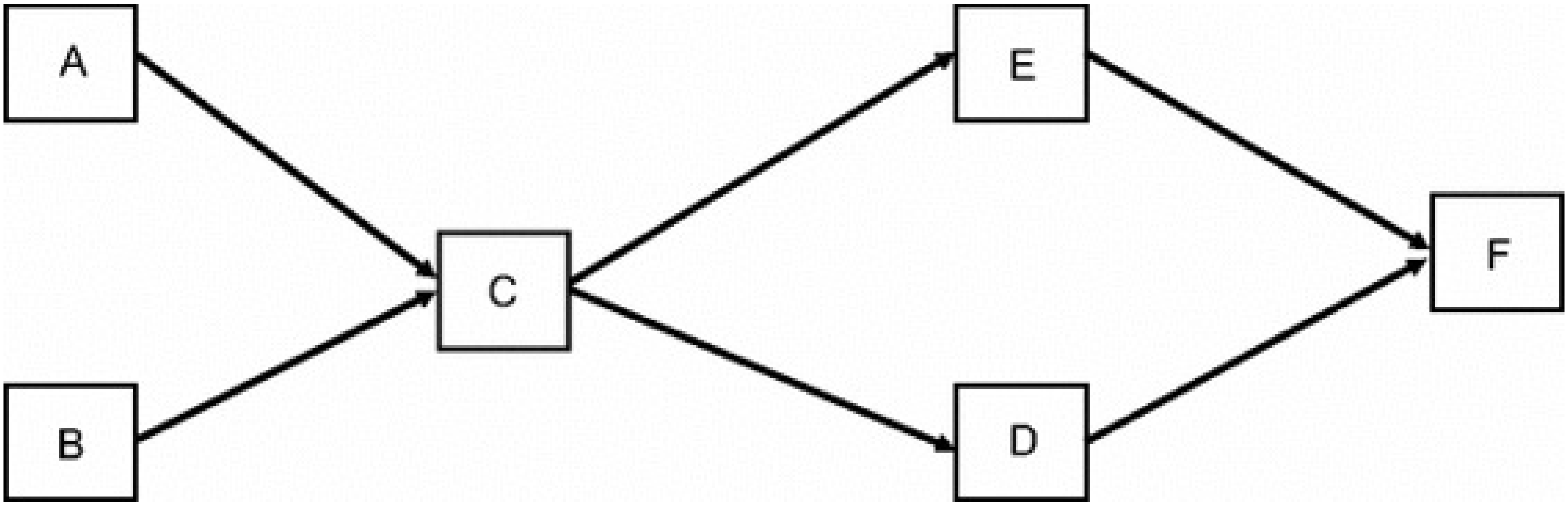


Figure 2.5 A directed graph describing the causal relationships between six variables or vertices (A to F)

Here is a *partial* verbal (as opposed to mathematical) description of what Figure 2.5 means. Two of the six variables (A and B) are *causally independent*, meaning that changes in either will not affect the value of the other. Each of the four other variables (C, D, E and F) are *causally dependent* on A and B, either directly (C) or indirectly (D, E and F). By 'causally dependent' I mean that changes in either A or B will provoke changes in each of C, D, E and F, but changes in any of these will not provoke changes in either A or B. A and B are *direct* causes of C because changes in A or B will provoke changes in C irrespective of the behaviour of either D, E or F. A and B are *indirect causes* of D, E and F because changes in A or B will provoke changes in these variables only by causing changes in C; if C is prevented from changing then A and B will no longer cause changes in these three other variables. C is a direct *common* cause of D and E and an indirect cause of F through its effects on D and E. Finally, both D and E are direct causes of F, though they are not themselves causally independent.

It is clear that this directed graph is a very compact way of expressing even the previous incomplete verbal description of this causal system. This economy of description is a major reason why researchers in artificial intelligence adopted directed graphs as a way of economically programming causal knowledge (Pearl 1988).[4] In order to better use and interpret directed graphs, a few definitions are needed.

In graph theory, a directed graph is a set of vertices, represented by letters enclosed in boxes in Figure 2.5, and a set of edges, represented by lines; these lines can have either no, single or double arrowheads. The arrowheads denote the direction of the functional relationship between the vertices at either end of the line.[5] Since biologists use directed graphs to represent causal relationships between variables, you can replace the abstract term 'vertex' with the more familiar word 'variable' and the abstract term 'edge' with the more familiar word 'effect'. The symbols at the ends of the lines can be either an arrowhead or a 'missing' mark. Thus, the notation X→Y means 'X is a direct cause of Y'. The notation X←Y means 'Y is a direct cause of X'. Finally, the notation X ↔ Y means 'neither X nor Y is a cause of the other but both share common unknown causes represented by some unknown vertex not included in the causal graph'. This last notation is needed later when we use incomplete causal graphs with unspecified latent vertices.

A *direct cause* is a causal relationship between two vertices that exists independently of any other vertex in the causal explanation. This is denoted by an arrow (→) whose tail is at the cause and whose head is pointing to its direct effect. For instance, both A and B are direct causes of C in Figure 2.5. Furthermore, A and B are the *causal parents* of C, and C is their *causal child*. A cause is 'direct' only in relation to the other vertices in the causal explanation. This point is important, because a common error is to incorrectly equate a 'direct' cause relative to others in the causal graph with the more fundamental claim that the cause is somehow 'direct' with respect to any other variable that might exist. Whenever you read the words 'direct cause' you should mentally add the words 'relative to the other variables that are explicitly invoked in the causal explanation'.

An *indirect cause* is a causal relationship between two vertices that is conditional on the behaviour of other vertices in the causal explanation. Again, a cause is 'indirect' only in relation to the other vertices in the causal explanation. For instance, in Figure 2.5 the vertex A is an indirect cause of vertex D (A→C→D) because its causal effect is conditional on the behaviour of vertex C. Furthermore, A and B are *causal ancestors* of D in Figure 2.5 and D is a *causal descendant* of both A and B.

Perhaps an example would help at this point. If we wish to give a causal description of the murder of a victim by a gunman and this explanation involves only these two 'variables' then we would say that the gunman's actions were the direct cause of the victim's death, and write 'Gunman's actions→Murder of victim'. On the other hand, if we also include the presence of the bullet penetrating the victim's heart in our causal explanation then we would say that the bullet was the direct cause of death and the gunman was an indirect cause, and write ‘Gunman's actions→Bullet→Murder of victim’. If we wish to go into more gruesome physiological detail then we would describe how the bullet interrupts the functioning of the heart and the bullet would no longer be a direct cause of the victim's death. Virtually any causal mechanism can be further decomposed into a more detailed causal mechanism, and so describing a cause as ‘direct’ or ‘indirect’ can be meaningful only in relative terms in the context of the other variables that make up the causal explanation. This is simply the reductionist method common in science, and the trick is always to choose a level of causal complexity that is sufficiently detailed that it meets the goals of the study while remaining applicable in practice.

A *directed path* between two vertices in a causal graph exists if it is possible to trace an ordered sequence of vertices that must be traversed, when following the direction of the edges (head to tail), in order to travel between the first and the second. If no such directed path exists then the two vertices *are causally independent*; causal conditional independence is defined below. It is possible for there to be more than one directed path linking two vertices. In Figure 2.5 there are two different directed paths between A and F: A→C→D→F and A→C→E→F.

An *undirected path* between two vertices in a causal graph exists if it is possible to trace an ordered sequence of vertices than must be traversed, *ignoring* the direction of the edges (head to tail), in order to travel from the first to the second. Be careful! An undirected path is *not* one having no arrowheads; rather, it is simply one resulting from ignoring them. In other words, an undirected path can also be a directed path, but this is not necessarily the case. For instance, there is an undirected path between A and B in Figure 2.5 (A→C←B) that is not also a directed path.

A *collider* vertex on an undirected path is a vertex with arrows pointing into it from both directions. The vertex F in the undirected path D→F←E in Figure 2.5 is a collider along this undirected path. It is possible for the same vertex to be a collider along one path and a non-collider along another path. A vertex that is a collider along an undirected path is *inactive* in its normal (unconditioned) state. This means that, in its normal (unconditioned) state, a collider blocks (prevents) the transmission of causal effects along such a path. The contrary of a collider is a *non-collider*. The vertex C in the path A→C→D in Figure 2.5 is a non-collider. A vertex that is a non-collider along a path is said to be active in its normal (unconditioned) state. This means that, in its normal (unconditioned) state, a non-collider permits the transmission of causal effects along such a path. It is sometimes easier to imagine a path as an electrical circuit and the variables (vertices) along the path as switches. A variable along a path that is a collider is like a switch that is normally OFF and a variable along a path that is a non-collider is like a switch that is normally ON.

An *unshielded collider* vertex is a set of three vertices A→B←C along a path such that B is a collider and, additionally, there is no edge between A and C. In Figure 2.5 the vertex F in the undirected path D→F←E is not only a collider but also an unshielded collider, since there is no edge between D and E. The contrary of an unshielded collider is a *shielded collider*.

2.3 Causal conditioning

I have been referring to the letters in the causal graph as ‘vertices’. Once we include the notion of a probability distribution that is generated by the causal graph, these vertices will also represent random variables. These vertices can be conceived to exist in one of two binary states along a given path: active or inactive. As stated above, the natural state of a non-collider is the active (ON) state and the natural state of a collider is the inactive (OFF) state. Again, it is possible for a vertex to be active along one path and inactive along another. Intuitively, one can think of the arrows as pointing out the direction of causal influence. Thus, a vertex that is both an effect and a cause (a type of non-collider), such as vertex C along the path A→C→D in Figure 2.5, is active, because it allows the causal influence of A to be transmitted to D. In the same way, a vertex that is an effect of two vertices and therefore a cause to neither (a collider) is inactive, because it blocks the causal influence from being transmitted along the path. An example is the vertex F along the path D→F←E in Figure 2.5. *Conditioning* on a vertex in a causal graph means to change its state; if it was active then conditioning inactivates it, but if it was inactive then conditioning activates it. So, since vertex C along the path A→C→D is naturally active (ON), conditioning on it changes its state to inactive (OFF), thus blocking any indirect causal influence of A on D.

2.4 D-separation

Remembering that we are still not discussing probability distributions or statistical models, and are still concerned only with the properties of directed acyclic graphs, we can now define what is meant by the 'independence' of vertices, or groups of vertices, in a causal graph upon conditioning on some other set of vertices. This property is called *d-separation* ('directed separation': Verma and Pearl 1988; Pearl 1988; Geiger, Verma and Pearl 1990). The definition of d-separation uses the definitions above and, although it is awkward to define in words, it is very easy to understand when looking at a causal graph. The formal definition is given in Box 2.1. I then give a more informal definition, and finally I illustrate it using figures.

Box 2.1 Formal definition of *d-separation*[6]

Given a causal graph G, if X and Y are two different vertices in G, and **W** is a set of vertices in G that does not contain X or Y, then X and Y are d-separated given **W** in G if and only if there exists no undirected path U between X and Y such that (a) every collider on U is either in **W** or else has a descendant in **W** and (b) no other vertex on U is in **W**.

Informally, d-separation gives the necessary and sufficient conditions for two vertices in a directed acyclic (causal) graph to be observationally (probabilistically) independent upon conditioning on some other set of vertices. Stated a different way, d-separation specifies how causal information about the effects of statistical conditioning on a causal system flows between vertices in the causal graph. Notice that I say causal *information* (what we can know about the result of statistical conditioning in a causal system) and not causal *effects*. D-separation is the translation device between the language of causality and the language of probability distributions that we have been searching for. In a few pages you will understand how this translation device works. Just be patient for a few more pages while I explain how to obtain d-separation claims from a causal graph.

To know if two vertices (X, Y) are d-separated given some set of other vertices in the causal graph, which we will call **W**, do the following.

(1) List every undirected path between X and Y. In other words, find every unique way that you can get to both X and Y in the causal graph if you ignore the direction of the arrowheads and then write down these paths.

(2) For every such undirected path between X and Y (which is an ordered sequence of vertices that must be traversed, ignoring the directions of the arrows), see if *any* non-colliding vertices in this path are in the conditioning set **W**. If so, then the path is blocked and there is no causal influence between X and Y along this path. Remembering that conditioning on a non-collider changes its state to inactive, at least one of the vertices in **W** blocks any causal influence between X and Y along this undirected path.

(3) For every such undirected path between X and Y, see if *every* collider vertex along this path is either a member of the conditioning set **W** or else has a causal descendant that is a member of the conditioning set **W**. If not, then the path is blocked and there is no causal influence between X and Y along this path. Remembering that conditioning on a collider changes its state from inactive to active, there is at least one collider along this undirected path that remains inactive, and so this path cannot transmit causal influence between X and Y.

The use of d-separation to deduce probabilistic independence upon conditioning from a causal system is best understood using a diagram (Figure 2.6) given by Spirtes, Glymour and Scheines (1993). First, we need some notation to speed things up. Whenever you see a d-separation (or 'd-sep') claim using the notation $X \perp\!\!\!\perp Y|\{\mathbf{W}\}$ you should read this as 'vertices X and Y are d-separated ("$\perp\!\!\!\perp$") given ("|") the set $\mathbf{W} = \{v_1, v_2, \ldots\}$ of conditioning vertices in the causal graph'. So, are vertices X and V d-separated in the causal graph in Figure 2.6 if we don't condition on any other vertices? Using our notation, and letting $\{\phi\}$ mean an empty set, is it true that $X \perp\!\!\!\perp V|\{\phi\}$? To get an answer, we first write out all undirected paths linking vertices X and V. There is only one, X→U←V. Since the causal information collides at U along this path, its natural (unconditioned) state is OFF and causal information cannot pass between X and V along this path. Next, we look at each non-collider vertex along each undirected path (there are none in our single path) and then check to see if any of these non-colliders are in our conditioning set (which is empty in this case). If we had found any non-colliders then our undirected path would be blocked, but this is not the case. Next, we look at each collider vertex along each undirected path (there is one: vertex U) along our single undirected path and check to see if all these colliders are in our conditioning set or else are ancestors of all the vertices in the conditioning set. In our case, our conditioning set is empty, and so our single undirected path (X→U←V) remained blocked. Therefore, X is d-separated from V if we don't condition on any variables. How about X ⫫ V|{U}? We have the same list of undirected paths as previously but now we condition on vertex U. Since U is a collider vertex along this undirected path, and U is also in the conditioning set, we change the status of this collider vertex from blocked (its natural state) to open. Therefore, information about the state of vertices X and V can flow between them when conditioning on vertex U. If this seems counter-intuitive to you then wait a few pages for the explanation.

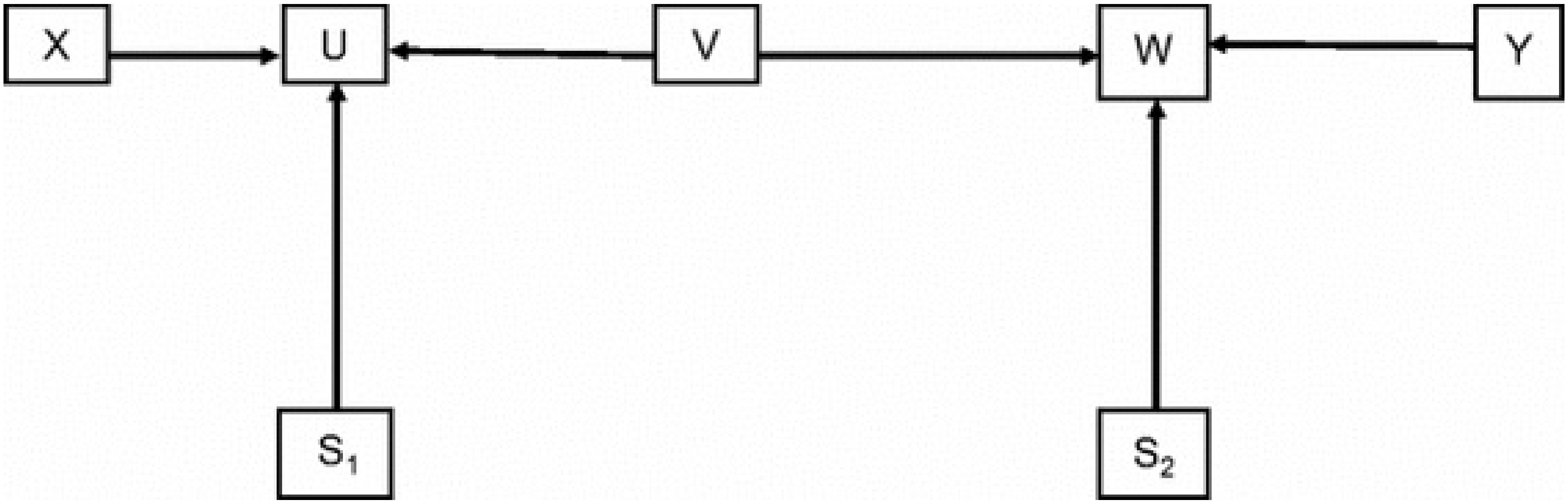


Figure 2.6 A directed graph used to illustrate the notion of d-separation

It is important to be able to use this d-sep operation when using causal graphs. Table 2.1 lists some of the d-separation statements that can be obtained from Figure 2.6; the negation ~X ⫫ Y|{W} means ‘vertices X and Y are *not* d-separated given the conditioning set **W**’.

Table 2.1 Various probabilistic independence relationships of the directed graph in Figure 2.6 that can be deduced using d-separation

Independence relation	Explanation
X ⫫ V|ϕ. X and V unconditionally independent.	There are no directed paths between X and V.
~X ⫫ V|U. X and V not independent, conditioned on U.	Since X→U←V collides at U, conditioning on U activates this path.
~X ⫫ V|S$_1$. X and V not independent, conditioned on S$_1$.	Since S$_1$ is a causal ancestor of U, conditioning on S$_1$ activates U along path X→U←V.

$\sim U \perp\!\!\!\perp W\|\phi$. U and W are not unconditionally independent.	The path U←V→W is naturally active. U and W share a common cause: (V) and V is not in the conditioning set $\{\phi\}$.
$U \perp\!\!\!\perp W\|V$. U and W are independent, conditioned on V.	There is only one naturally active path between U and W: U←V→W. Conditioning on V inactivates V, blocking this path.
$X \perp\!\!\!\perp Y\|\phi$. X and Y are unconditionally independent.	The only undirected path between X and Y is naturally blocked by both U and W.
$\sim X \perp\!\!\!\perp Y\|\{U,W\}$. X is not independent of Y, conditioned simultaneously on U and W.	The only undirected path between X and Y has two colliders, and both are in the conditioning set. This activates the undirected path.
$\sim X \perp\!\!\!\perp Y\|\{S_1,S_2\}$. X is not independent of Y, conditioned simultaneously on S_1 and S_2.	The only undirected path between X and Y has two colliders, and the causal ancestors of both are in the conditioning set. This activates the undirected path.
$X \perp\!\!\!\perp Y\|\{U,W,V\}$. X is independent of Y, conditioned simultaneously on U, W and V.	Although conditioning on both U and W activates these two colliders, conditioning on V disactivates this non-collider.

The causal inferences about this graph that are listed in Table 2.1 are not exhaustive. After a few minutes of practice it is easy to simply read off the conditional independence relations from such a causal graph. However, there also exists a function (dSep) in the ggm (graphical Gaussian model) library of R that can give the answers to d-sep claims.

The first step (after loading the ggm library) is to input your DAG. This is done using the DAG function and the ~ operator of R that is commonly used for specifying model formulae. The syntax of the DAG function is DAG(…, order = FALSE). The first argument (…) is the sequence of model formulae that specifies the DAG. The second argument (order) specifies if you want to keep the model formulae in the order in which you entered them (order = FALSE), which is the default, or if you want the order rearranged to respect the topological of cause–effect ordering. Using Figure 2.6 as an example, here is how you would use the DAG function to specify this causal graph:

```
Figure. 2.6←DAG(
U~X+V,
S1~U,
W~V+Y,
S2~W,order = FALSE)
```

There is some flexibility in the specification of the DAG. You can enter each parent–child node separately. For instance, rather than entering U~X+Y you could enter U~X, U~Y. If you have a vertex (say 'v') that is isolated (i.e. one that neither causes nor is caused by another variable) then you would specify this as v~v. The output of the DAG function is a square Boolean matrix in which the rows represent the causal parents, the columns represent the causal children and a '1' in cell (i,j) means that there is an arrow going from row i to column j. The row names attribute of this matrix are the names that you input for your vertices.

After you have input your causal graph using the DAG function and saved it, then you can input d-sep queries using the dSep function. The syntax of the dSep function is

```
dSep(amat,first,second,cond)
```

. The first argument, 'amat', is the Boolean matrix that you would normally create using the DAG function. The next three arguments are the names of the two vertices whose d-separation status you wish to know ('first' and 'second') and the names of the conditioning vertices ('cond'). Each of these last three arguments can be specified as single names or as a vector of names. The name of an empty set is 'NULL'.

The first d-sep claim in Table 2.1 is $X \perp\!\!\!\perp V|\{\phi\}$. Here is how to use the dSep function to ask if X is d-separated from V in Figure 2.6 if no other vertices are fixed:

```
dSep(Figure. 2.6,first = "X",second = "V",cond = NULL)
[1] TRUE
How about XY|{U,W}?
dSep(Figure. 2.6,first = "X",second = "Y",cond =
c("U","W"))
[1] FALSE
```

D-separation leads to a wealth of very useful results involving causal inference, many of which will be described in later chapters. However, until d-separation is related to probability distributions it provides no way of inferring causal relationships from observational data. Before making this link explicit, we first need some notions from probability theory.

2.5 Probability distributions

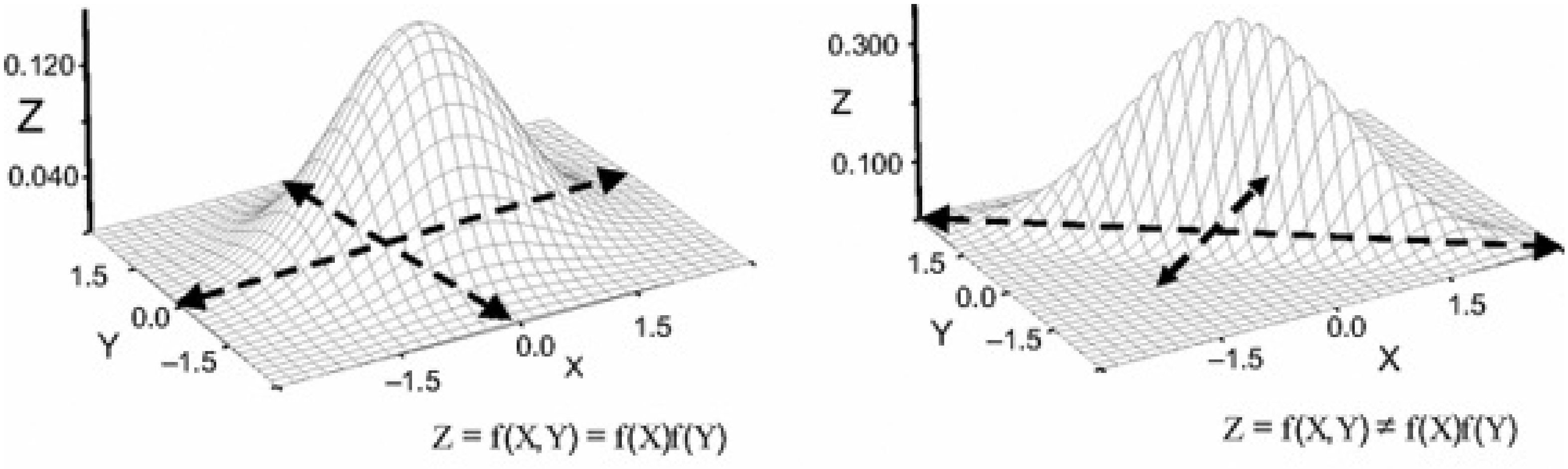
The vertices of a causal graph represent attributes in a causal system (e.g. the nitrogen concentration in a leaf or the body mass of a sheep). When we randomly sample observational units (leaves, sheep) possessing these attributes (nitrogen concentration, body mass) from some statistical population that is governed by this causal system, the vertices of the causal graph are also random variables that obey a probability distribution. Since causal relationships involve at least two such random variables, we must deal with joint probability distributions.

As I have already briefly mentioned, the notion of ‘probability’ differs depending on whether one subscribes to a frequentist, objective Bayesian or subjective Bayesian school of statistics. Since most statistical methods familiar to biologists derive from a frequentist perspective, I will use this definition. One begins with a hypothetical statistical population (say, all wheat plants grown in Europe) that contains all the observational units (individual plants) of interest. Each observational unit has a variable (say, the protein content of a seed) that can take different values (1.2 mg, 3.1 mg, etc.). The proportion of observational units (individual plants) in the statistical population (wheat grown in Europe) taking different values of the variable of interest (seed protein content) is the probability of this variable in this statistical population. Another way of saying this is that the probability of a random variable (X) taking a value $X = x_i$ (or having a value within an infinitesimal interval around x_i) in a statistical population of size N is the limiting frequency of $X = x_i$ in a random sample of size n as n approaches N.

A probability distribution is the distribution of the limiting (relative) frequencies of $X = x_1, x_2, \ldots$in such a statistical population. Happily, it is an empirical fact that the distribution of many variables, when randomly sampled, can be closely approximated by various mathematical functions. Many of these functions are well known to biologists (normal distribution, Poisson distribution, binomial distribution, Fisher's F-distribution, chi-square distribution), and there are many less well-known functions that can be used as well. It is always an empirical question whether or not one of these mathematical distributions is a sufficiently close approximation of one's data to be acceptable. For instance, the relative frequency of the seed protein content per plant is likely to follow a normal distribution. The formula for the normal distribution is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

When only one variable is measured on each observational unit then one obtains a *univariate* distribution. When one measures more than one variable on each observational unit (say, both the protein content and the average seed weight per plant) then one obtains a *multivariate* distribution.[7] If one obtains the relative frequencies of values of each unique set of multivariate observations then one has a multivariate probability distribution. Again, there are many multivariate mathematical functions that approximate such multivariate probability distributions. Figure 2.7 shows two versions of a bivariate normal distribution.



(a) The joint distribution of two independent, normally distributed random variables

(b) The joint distribution of two normally distributed random variables that are not independent

Figure 2.7 Two different versions of a bivariate normal probability distribution

2.6 Probabilistic (conditional) independence

By definition, two random variables (X, Y) are (unconditionally) independent if the joint probability density of X and Y is the product of the probability density of X and the probability density of Y. Let's use the notation I(X,Φ,Y) to mean that random variables X and Y are independent, conditional on no other variables (i.e. our empty set Φ). Thus:

$$\text{if I(X, } \phi \text{, Y) then } P(X, Y) = P(X) \cdot P(Y)$$

For instance, if X and Y are each distributed as a standard normal distribution and they are also independent (Figure 2.7(a)) then the joint probability distribution can be obtained as follows:

$$f(X; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(X)^2}{2}}$$

$$f(Y; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(Y)^2}{2}}$$

$$f(X, Y) = f(X; 0, 1) \cdot f(Y; 0, 1) = \frac{1}{2\pi} e^{\frac{-(X^2+Y^2)}{2}}$$

If two random variables (X, Y) are not (unconditionally) independent then the joint probability density of X and Y is not the product of the two univariate probability densities. If the variables are dependent then one cannot simply multiply one univariate probability density to the other, because we have to take into consideration the interaction of the two (Figure 2.7(a)).

Figure 2.7(a) shows the bivariate normal density function of two independent variables. Note that the mean value of Y is the same (zero) no matter what the value of X, and vice versa; the value of one variable doesn't change with changes in the average, or *expected*, value of the other variable. Knowing that the value of X is 1.5 rather than –1.5 in Figure 2.7(a) doesn't give us any additional information about the values of Y that we will encounter. Figure 2.7(b) shows the bivariate normal density function of two dependent variables. Here, the mean value of Y is not independent of the value of X because, as the value of X increases, the mean value of Y decreases. Now, knowing that the value of X is 1.5 rather than –1.5 tells us that the value of Y is more likely to be closer to –1.5 than to 1.5.

Similarly, X and Y are independent, conditional on ('given') a set of other variables **Z**, if the joint probability density of X and Y given **Z** equals the product of the probability density of X given **Z** and the probability density of Y given **Z** for all values of X, Y and **Z** for which the probability density of **Z** is not equal to zero.[8] The notion of conditional independence will be explained in more detail in Chapter 3. Thus:

if I(X,**Z**, Y) then $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$

Because the notion of conditional independence is sometimes difficult to grasp, here is a verbal description. To say that a variable X is a random variable obeying a particular probability distribution is to say that we do not know for certain what the value of this variable will be when we next measure it on an observational unit but that we do have information about how likely it is to take on different values. What does it mean to say that our random variable X is dependent on some other random variable Y, also obeying some particular probability distribution? This means that we will have even more information about the likely values of X if we are given the value of Y ('conditional on Y') than if we don't know the value of Y. What if our random variable X is dependent on random variable Y, but conditionally independent given the value of some third random variable, Z? This means that whatever information we have about X that is provided by Y (and vice versa) is also provided by Z, and so, once we know Z, Y provides no *additional* information.

Let's make this a bit less abstract. Imagine an electrical circuit in which two light bulbs are each controlled by the same ON/OFF switch (Figure 2.8). Each bulb can also be turned on or off by its own separate switch. We can see the light bulbs but can't see the switches. For each bulb, if either switch (its own unique one or the common one) is on, the bulb lights up. These switches turn ON and OFF at random intervals and are not connected together. Is the light bulb on the right lighted or dark? If I tell you that the light bulb on the left is alight, does this information make it more or less likely that the light bulb on the right is also alight? Since one cause of the left bulb being turned on is that the common switch is ON, knowing that the left bulb is alight increases the *chances* that the bulb on the right is also ON. This means that the two random variables describing the states of the two light bulbs are dependent random variables. Similarly, if I tell you only that the common switch is on, you also know something extra about the state of the bulb on the right: it is certainly alight. This means that these two random variables (the states of the common switch and the second light bulb) are also dependent random variables. Now, if I tell you not only that the common switch is on (information about this random variable) but also that the light bulb on the left is on (information about this second random variable), this second bit of information about the state of the left light bulb gives you no extra information about the state of the right light bulb (the third random variable) that was not already provided by the first bit of information (about the common switch). In other words, the two random variables referring to the two light bulbs are independent (provide no mutual information) given ('conditional on') the information contained in the third random variable.

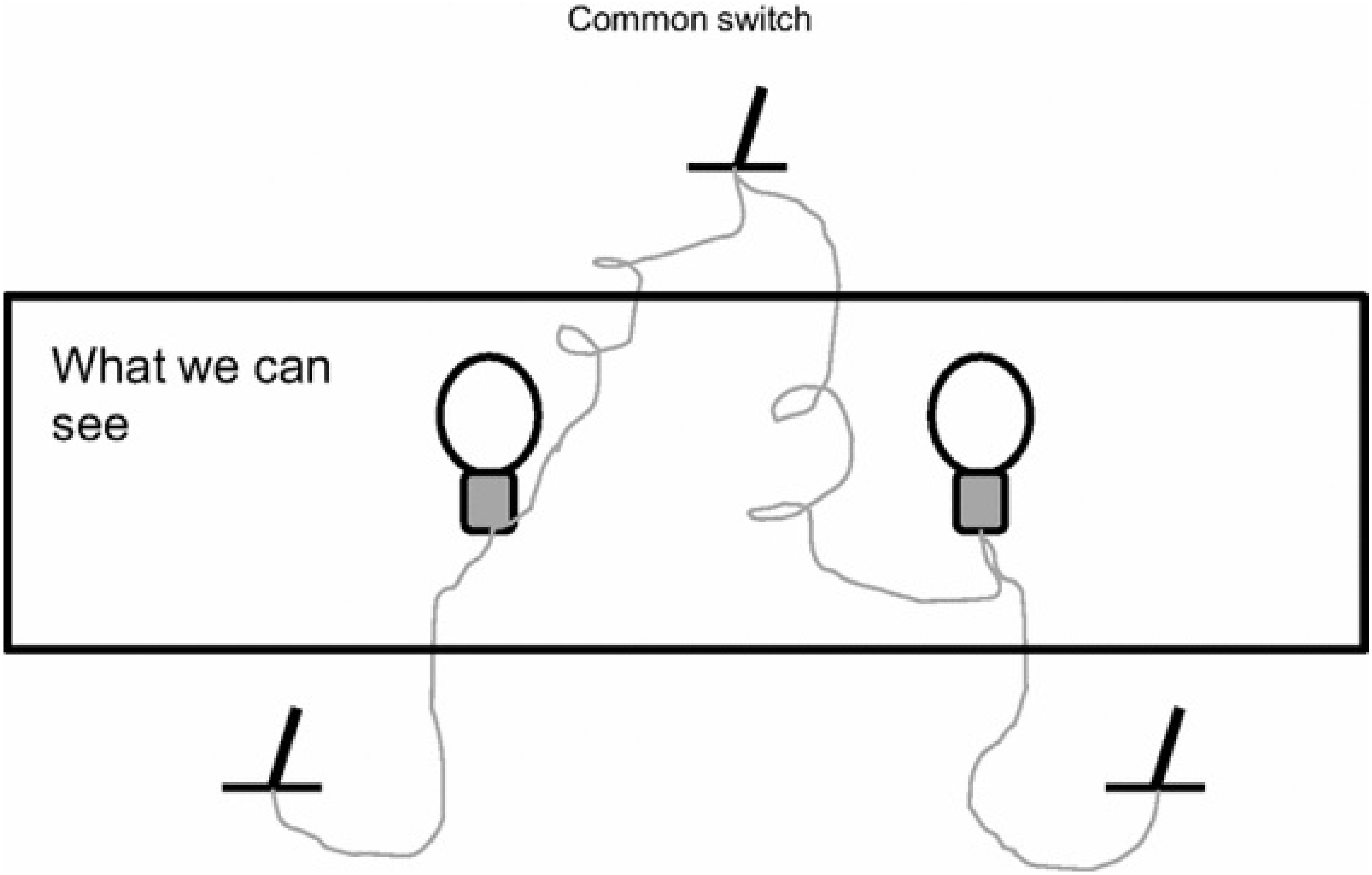


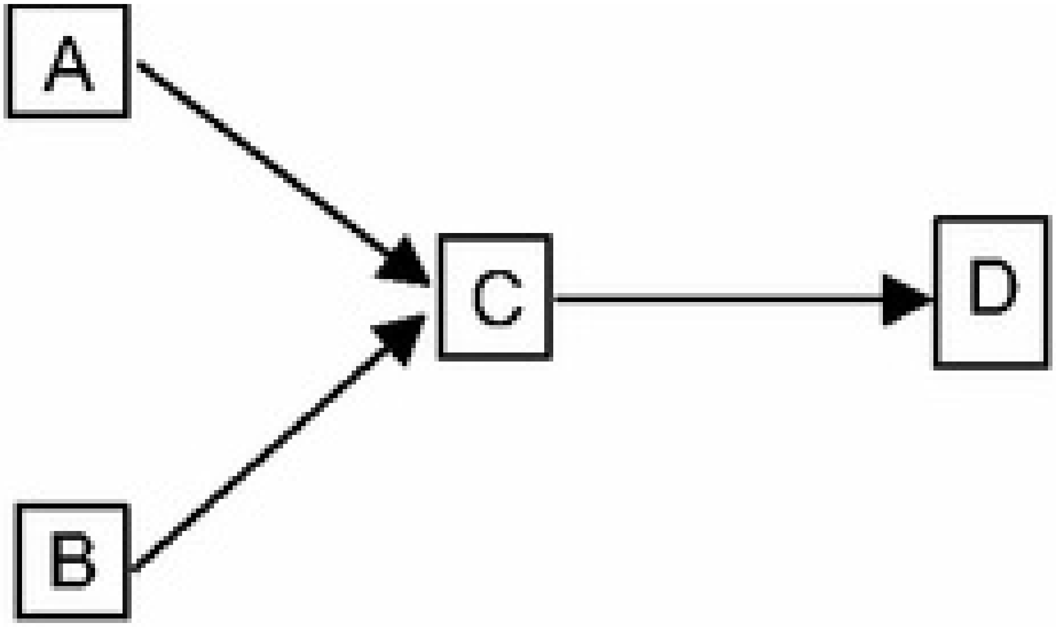
Figure 2.8 An electrical circuit in which two light bulbs are each controlled by the same ON/OFF switch; each bulb can also be turned ON independently with its own unique switch

2.7 The Markov condition

Many ecologists, especially those who study vegetation dynamics, are familiar with Markov chain models (van Hulst 1979). These models predict vegetation dynamics based on a 'transition matrix'. The transition matrix gives the probability that a location that is occupied by a species s_i at time t will be replaced by species s_j at time t+1. The model is 'Markovian' because of the assumption that changes in the vegetation at time t+1 depend at most on the state of the vegetation at time t, but not on states of the vegetation at earlier times. Stated another way, these models are Markovian because they assume that the more distant past (t–1) affects the immediate future (t+1) only indirectly through the present (t); thus, (t–1)→(t)→(t+1). Stated a third way, these models are Markovian because they assume that the state of the random variable representing the immediate future is independent of the more distant past, conditional on the present state.

In the context of causal models, the Markov condition is a property both of a directed acyclic (causal) graph and the joint probability distribution that is generated by the graph. The condition is satisfied if, given a vertex v_i in the graph, or a random variable v_i in the probability distribution, v_i is independent of all ancestral causes given its causal parents.[9] In the context of a causal model, this assumption is simply the claim that, once we know the direct causes of an event, knowledge of more distant (indirect) causes provides no new information. Notice that this is one of the properties stipulated in Chapter 1 for our mathematical language of causality. To use a previous example,[10] assume that the only cause of an increased concentration of photosynthetic enzymes in a leaf is the added fertiliser that was put on the ground, and that the only cause of an increased photosynthetic rate is the increased concentration of photosynthetic enzymes. Then, knowing how much fertiliser was added gives us no new information about the photosynthetic rate once we already know the concentration of photosynthetic enzymes in the leaf.

An important property of probability distributions that obey the Markov condition is that they can be decomposed into conditional probabilities involving only variables and their causal parents. For example, Figure 2.9 shows a causal graph and the joint probability distribution that is generated by it. This decomposition states that, to know the probability distribution of D, we need only to know the value of C – i.e. P(D|C). To know the probability distribution of C we need only to know the values of A and B – i.e. P(C|{A,B}). A and B are independent, and so to know the joint probability distribution of A and B we need only to know the marginal distributions of A and B – i.e. P(A)P(B).



$$P(A, B, C, D) = P(A) \bullet P(B) \bullet P(C \mid \{A, B\}) \bullet P(D \mid C)$$

Figure 2.9 A causal graph involving four variables and the joint probability distribution that is generated by it

2.8 The translation from causal models to observational models

Although causal models and observational models are not the same thing, there is a remarkable relationship between the two. Consider the case of causal graphs that do not have feedback relationships – that is, directed paths from some vertex that do not lead back to the same vertex. Theorem 10 by Pearl (1988) states that, for any causal graph without feedback loops (a directed acyclic graph), every d-separation statement obtained from the graph implies an independence relation in the joint probability distribution of the random variables represented by its vertices.

This central insight has been a long time in coming, and I imagine that many readers will wonder whether the effort was worth the return, so let me rephrase it:

> Once we have specified the acyclic causal graph, then every d-separation relation that exists in our causal graph must be mirrored in an equivalent statistical independency in the observational data if the causal model is correct.

The above statement is incredibly general. It does not depend on any distributional assumptions of the random variables or on the functional form of the causal relationships. In the same way, if even one statistical independency in the data disagrees with what d-separations of the causal graph predict, the causal model must be wrong. This is the translation device that we needed in order to properly translate the causal claims represented in the directed graph into the 'official' language of probability theory used by statisticians to express observational models. After wading through the jargon developed above, I hope that you will recognise the elegant simplicity of this strategy (Figure 2.10). First, express one's causal hypothesis in a mathematical language (directed graphs) that can properly express the asymmetric types of relationships that scientists imply when they use the language of causality. Second, use the translation device (d-separation) to translate from this directed graph into the well-known mathematical language (probability theory) that is used in statistics to express notions of association. Finally, determine the types of (conditional) independence relationships that must occur in the resulting joint probability distribution. The hidden causal processes cast observational shadows in nature's shadow play. D-separation is the translation device by which one can predict the shape of these shadows given an hypothesised causal process. The shadows are in the form of conditional independence relationships that the joint probability distribution (and therefore the observational model) must possess if the data are really generated by the hypothesised directed graph.

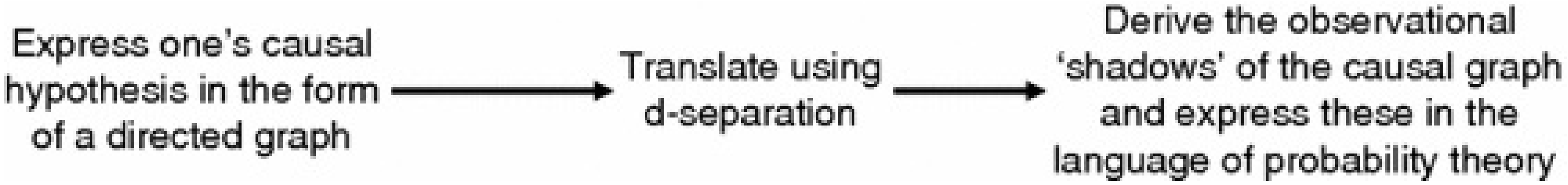

Express one's causal hypothesis in the form of a directed graph
Translate using d-separation
Derive the observational 'shadows' of the causal graph and express these in the language of probability theory

Figure 2.10 The strategy used to translate from a causal model to an observational model

2.9 Counter-intuitive consequences and limitations of d-separation: conditioning on a causal child

D-separation does not tell us how a causal system will respond following an external manipulation.[11] Rather, d-separation is a mathematical operation that gives the correlational *consequences* of conditioning on a variable in a causal system. One non-intuitive consequence is that two causally independent variables will be correlated if one conditions on any of their common descendants. This is because conditioning on a collider vertex along a path between vertices X and Y means that X and Y are not d-separated. This has important consequences for applied regression analysis and shows how such a method can give very misleading results if these are interpreted as giving information about causal relationships.

Consider a causal system in which two causally independent variables (X and Y) jointly cause variable Z: X→Z←Y. To be more specific, let's assume that the nitrogen content (X) and the stomatal density (Y) of the leaves of individuals of a particular species jointly cause the observed net photosynthetic rate (Z). Further, assume that leaf nitrogen content and stomatal density are causally independent. The causal graph is therefore leaf nitrogen→net photosynthetic rate←stomatal density. Let the functional relationships between these variables be as follows:

$$
\begin{aligned}
&\mathit{leaf\ nitrogen} = N(0, 1)\\
&\mathit{stomatal\ density} = N(0, 1)\\
&\mathit{net\ photosynthesis} = 0.5\,\mathit{leaf\ nitrogen} + 0.5\,\mathit{stomatal\ density} + N(0, 0.707)
\end{aligned}
$$

These three equations can be used to conduct numerical simulations[12] that can demonstrate the consequences of conditioning on a common causal child (the net photosynthetic rate). Since I will use this method repeatedly in this book, I will explain how it is done in some detail. The first equation states that the leaf nitrogen concentration of a particular plant has causes not included in the model. Since the plant is chosen at random the leaf nitrogen concentration is simulated by choosing at random from a normal distribution whose population mean is zero and whose population standard deviation is 1. The second equation states that the stomatal density of the same leaf of this individual also has causes not included in the model (not the same unknown causes, since otherwise it would not be causally independent) and its value is simulated by choosing another (independent) number from the same probability distribution. The third equation states that the net photosynthetic rate of this same leaf is jointly caused by the two previous variables. The quantitative effect of these two causes on the net photosynthetic rate is obtained by adding 0.5 times the leaf nitrogen concentration plus 0.5 times the stomatal density plus a new (independent) random number taken from a normal distribution whose population mean is zero, whose population variance is 1–2(0.5^2) and whose population standard deviation is therefore the square root of this value; this third random variable represents all those other causes of net photosynthetic rate other than leaf nitrogen and stomatal density, and these other unspecified causes are not causally connected to either of the specified causes.

By repeating this process a large number of times, one obtains a random 'sample' of 'observations' that agree with the generating process specified by the equations.[13] As will described in Chapter 3, this model is actually a very simple path model. Here is some simple R code to do this simulation:

```
leaf.nitrogen←rnorm(1000,0,1)
stomatal.density←rnorm(1000,0,1)
net.photosynthesis←0.5*leaf.nitrogen+0.5*stomatal.density-

rnorm(1000,0,0.707)
```

After generating 1,000 independent 'observations' that agree with these equations, and respecting the causal relationships specified by our causal system, these are the regression equations that are obtained:

$$\begin{aligned}
&\textit{leaf nitrogen} = N(0.002, 1.001)\\
&\textit{stomatal density} = N(-0.044, 1.000)\\
&\textit{net photosynthesis} = 0.001 + 0.470\ \textit{leaf nitrogen} + 0.514\ \textit{stomatal density} + N(0, 0.707)
\end{aligned}$$

Happily, the partial regression coefficients as well as the means and standard deviations of the random variables are what we should find, given sampling variation with a sample size of 1,000.

What happens if we give these data to a friend who mistakenly thinks that leaf nitrogen concentration is actually caused by net photosynthetic rate and stomatal density? In other words, she mistakenly thinks that the causal graph is net photosynthetic rate→leaf nitrogen←stomatal density. We know, because we generated the numbers, that leaf nitrogen and stomatal density are actually independent (the Pearson correlation coefficient between them is –0.037, which is not statistically significant at the traditional 0.05 confidence level), but this is the set of regression equations that results from this incorrect causal hypothesis:

$$\begin{aligned}&\textit{net photosynthesis} = N(-0.002, 0.987)\\&\textit{stomatal density} = N(-0.044, 1.000)\\&\textit{leaf nitrogen} = 0.000 + 0.654\ \textit{net photosynthesis} - 0.350\ \textit{stomatal density} + N(0, 0.834)\end{aligned}$$

Tests of significance for the two partial regression coefficients show that each is significantly different from zero at a probability of less than 1×10^{-6}. Why would the multiple regression mistakenly report a highly significant ‘effect’ of stomatal density on leaf nitrogen when we know that they are both statistically (remember that the correlation between the two was only –0.037) and causally independent (because we made them that way in the simulation)? There is no ‘mistake’ in the statistics; rather, it is due to our mistranslation between the language of probability and the language of causality. The regression equation is an observational model. It is simply telling us that knowing something about the stomatal density gives us extra information about (or helps to predict) the amount of nitrogen in the leaf, *when we compare leaves with the same net photosynthetic rate*.[14] This is exactly what d-separation, applied to the correct causal graph, tells us will happen: leaf nitrogen and stomatal density, while unconditionally d-separated, are not d-separated (therefore observationally associated) upon conditioning on their causal child (the net photosynthetic rate).

This counter-intuitive claim is easier to understand with an everyday example. Consider again the simple causal world consisting only of rain, watering cans and mud, related as rain→mud←watering cans. Now, in this world there are no causal links between watering cans and rain. Knowing that no one has dumped water from the watering can tells us nothing about whether or not it is raining; we can predict nothing about the occurrence of rain by knowing something about the watering can. On the other hand, if we see that there is mud (the causal child of the two independent causes) *and* we know that no one has dumped water from the watering can (i.e. conditional on this variable) then we can predict that it has rained. Conditioning on a common child of the two causally independent variables (rain and watering cans) renders them observationally dependent. This is because information, unlike causality, is symmetric.

Many researchers believe that the more variables that can be statistically controlled in a multiple regression, the less biased and the more reliable the resulting model. The above example shows this to be wrong and warns against such methods as stepwise multiple regression if the resulting model is to be interpreted as something more than simply a prediction device.[15] This point is rarely mentioned in most statistics texts.

2.10 Counter-intuitive consequences and limitations of d-separation: conditioning due to selection bias

There is also an interesting consequence of d-separation that might occur in selection experiments. ‘Body condition’ is a somewhat vague concept that is sometimes used to refer to the general health and vigour of an animal. It is occasionally operationalised as an index based on a weighting of such things as the amount of subcutaneous fat, the parasite load or other variables judged relevant to the health of the species. Imagine a wildlife manager who wants to select for an improved body condition of bighorn sheep. His measure of body condition is obtained by adding together the thickness of subcutaneous fat in the autumn (cm) and a score for parasite load (0 = none, 1 = average load, 2 = above-average load) as follows: body condition = 0.5fat + parasite load. These two components of body condition are causally unrelated. He decides to protect all individuals whose body condition is greater than 3 and removes all others from the population by allowing hunters to kill them. The causal graph of this process is fat thickness→body condition←parasite load. If someone else were to then measure the fat thickness and parasite load in the remaining population after the selective hunt, she would find that these two variables were correlated, even though there is, in reality, no causal link between the two.[16] This occurs because the selective hunt has removed all those individuals not meeting the selection criterion, and this effectively results in conditioning on body condition.

2.11 Counter-intuitive consequences and limitations of d-separation: feedback loops and cyclic causal graphs

The relationship between d-separation in an acyclic causal model (a DAG) and independencies in a probability distribution is therefore very general. What happens if there are feedback loops in the causal model? We don't know for sure, although this is an area of active research (Richardson 1996b). Spirtes (1995) has shown that d-separation in a cyclic causal model still implies independence in the joint probability distribution that it generates, but only if the relationships are linear. Pearl and Dechter (1996) have also shown that the relationship between d-separation and probabilistic independence also holds if all variables are discrete without any restriction on the functional form of the relationships. Unfortunately, Spirtes (1995) has also shown, by a counter-example, that d-separation does not always imply probabilistic independence when the functional relationships are non-linear and the variables are continuous. There are some grammatical constructs in the language of causality for which no one has yet found a good translation.

There are other curious properties of causal models with feedback loops. Consider Figure 2.11: such a causal model seems to violate many properties of causes. The relationship is no longer asymmetric, since X causes Z (indirectly through Y) and Z also causes X. The relationship is no longer irreflexive, since X seems to cause itself through its effects on Y and Z.

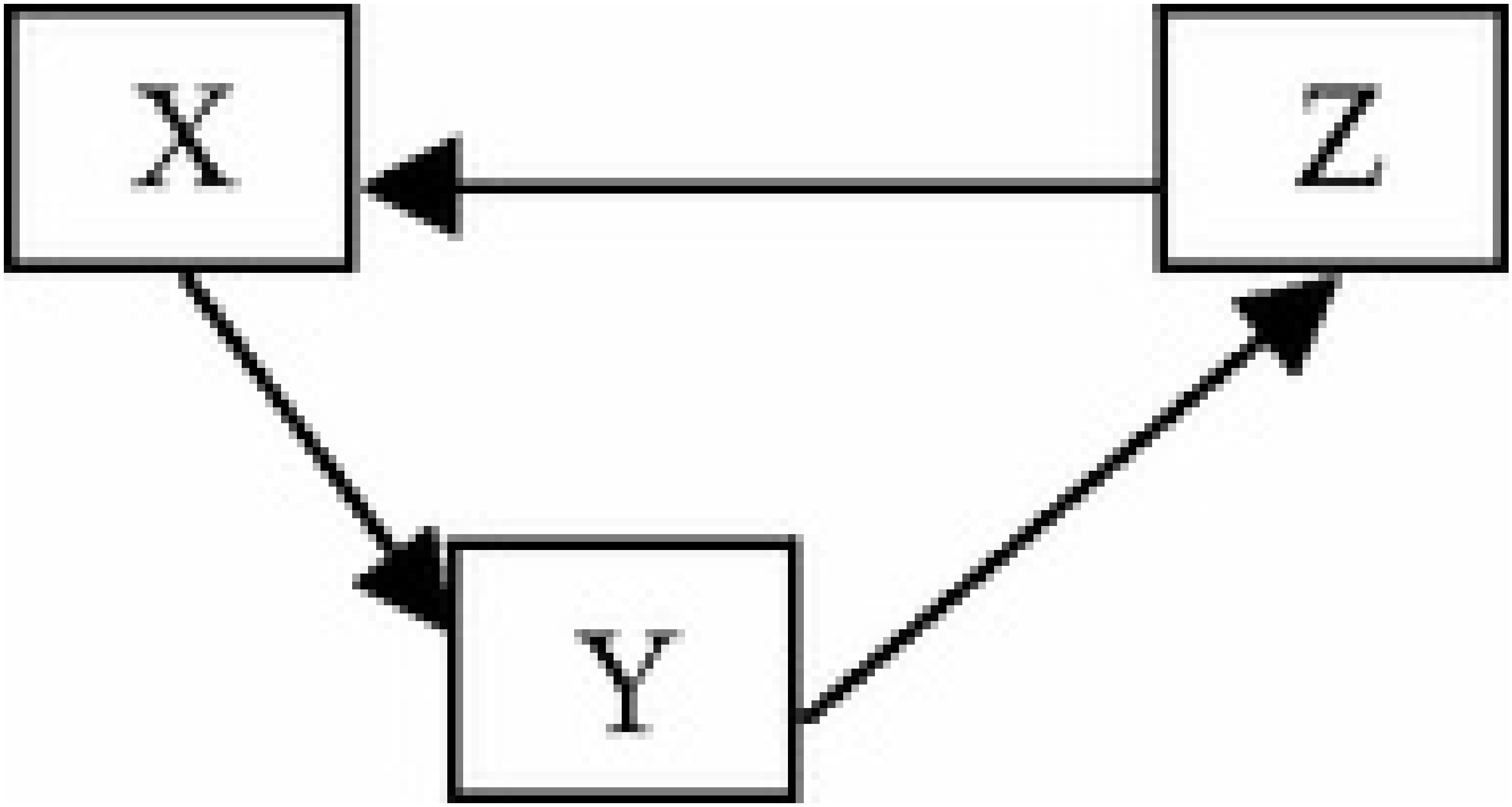


Figure 2.11 A cyclic causal graph that seemingly violates many of the properties of 'causal' relationships

These counter-intuitive aspects of feedback loops can be resolved if we remember that causality is a process that must follow time's arrow but causal graphs do not explicitly include this time dimension. Causal graphs with feedback loops represent either a 'time slice' of an ongoing dynamic process or a description of this dynamic process at equilibrium, an interpretation that appears to have been first proposed by F. M. Fisher (1970). Thomas Richardson's very interesting PhD thesis (Richardson 1996b) provides a history of the use and interpretation of such cyclic, or 'feedback', models[17] in economics. A more complete causal description of the process shown in Figure 2.11 is given in Figure 2.12; the subscripts on the vertices index the state of that vertex at a given time. From Figure 2.12 we see that, once the explicit time dimension is included in the directed graph, the apparent paradoxes disappear. Rather than circles, when we ignore the time dimension (as in Figure 2.11), we have spirals that never close on themselves when the time dimension is included. Just as the one-year-old Bill Shipley is not the same individual as I am as I write these words, the 'X' that causes Y at time t = 1 will not be that same 'X' that is caused by Z at time t = 4 in Figure 2.12.

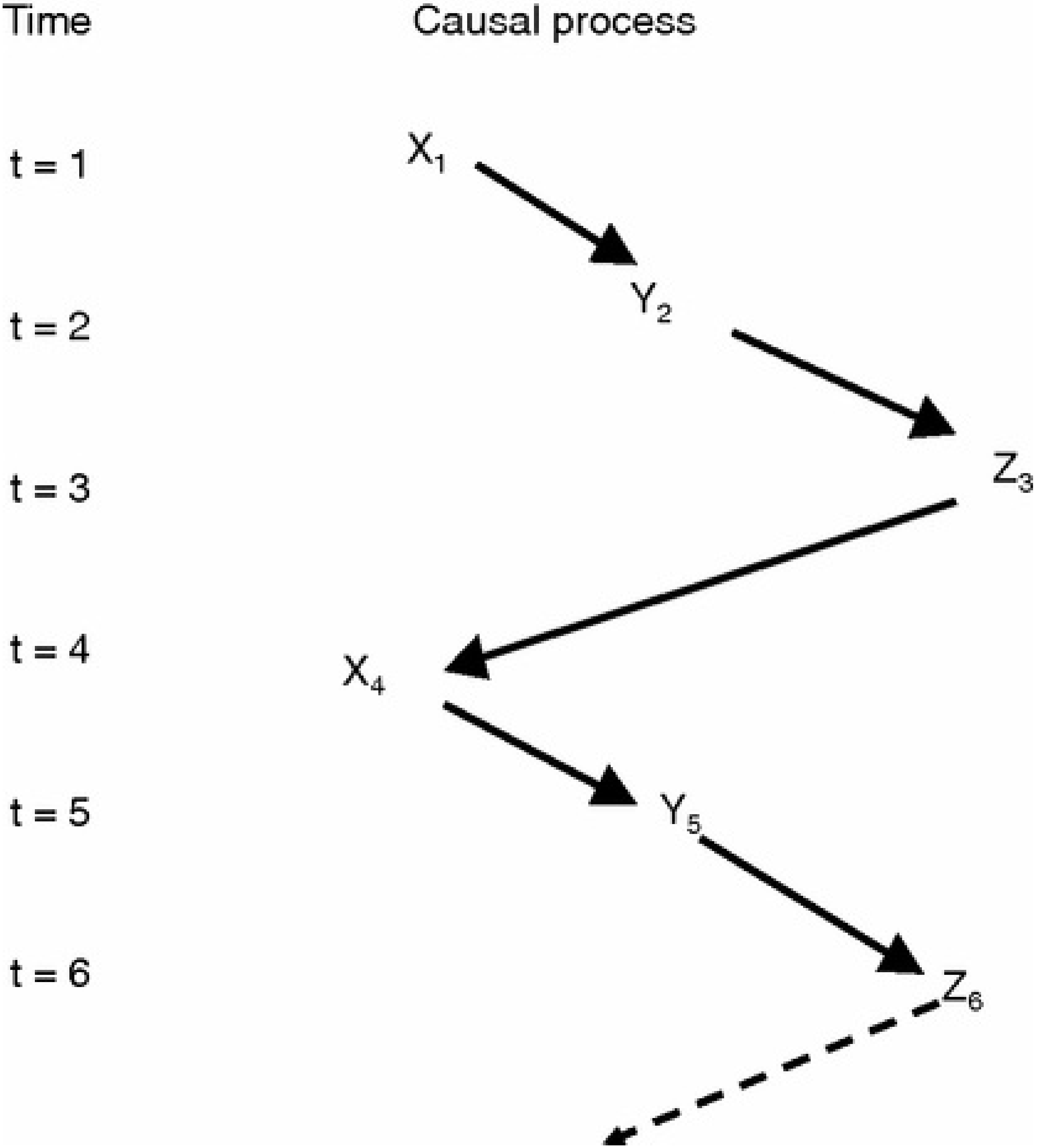


Figure 2.12 The causal relationships between X, Y and Z from Figure 2.11 when the time dimension is included in the causal graph

Conceived in this way, both acyclic and cyclic causal models represent ‘time slices’ of some causal process. Samuel Mason, described by Heise (1975), provided a general treatment of feedback loops in causal graphs over 60 years ago for the case of linear relationships between variables. Nonetheless, trying to model causal processes with feedback using directed graphs that ignore this time dimension is more complicated and requires one to make assumptions about the linearity of the functional relationships.

2.12 Counter-intuitive consequences and limitations of d-separation: imposed conservation relationships

Relationships derived from imposed (as opposed to dynamic) conservation constraints are superficially similar to cyclic relationships, but they are conceptually quite different. By ‘conservation’ I mean variables that are constrained to maintain some conserved property. For instance, if I purchase fruits and vegetables in a store and then count the total amount of money that I have spent, I can represent this as money spent on fruits→total money spend←money spent on vegetables. If the total amount of money that I can spend is not fixed then the amount that I spend on fruits and the amount that I spend on vegetables are causally independent. However, if the total amount of money is fixed, or *conserved*, due to some influence outside the causal system then every dollar that I spend on fruit causes a decrease in the amount of money that I spend on vegetables. There is now a causal link between the amount of money spent on fruits and on vegetables due only to the requirement that the total amount of money be conserved.

There is no obvious way to express such relationships in a causal graph. One might be tempted to modify our original acyclic graph by adding a cyclic path between ‘Fruits’ and ‘Vegetables’ but, if we do this, one cannot interpret such a cyclic graph as a static graph of a dynamic process; the conservation constraint is imposed from outside and is not due to a dynamic equilibrium that results from the prior interaction of ‘Money spent on fruits’ and ‘Money spent on vegetables’. In other words, it is not as if I spend one dollar more on fruits at time t = 1, which causes me to spend one dollar less on vegetables at time t = 2, which then causes me to spend one dollar less on fruits at time t = 3, and so on until some dynamic equilibrium is attained. The conservation of the total amount of money spent is imposed from outside the causal system.

One might also be tempted to interpret the conservation requirement as equivalent to physically fixing the total amount of money at a constant value. If this were true then one could maintain the causal graph ‘Money spent on fruits→Total money spent←Money spent on vegetables’ but with the variable ‘Total money spent’ being fixed due to the imposed conservation requirement. Because ‘Total money spent’ is now viewed as being fixed rather than being allowed to randomly vary, ‘Money spent on fruits’ would not be d-separated from ‘Money spent on vegetables’ (remember d-separation); this is because ‘Total money spent’ is the causal child of each of ‘Money spent on fruits’ and ‘Money spent on vegetables’. This would indeed imply a correlation between ‘Fruits’ and

‘Vegetables’. Unfortunately, our causal system does not simply imply that the money spent on fruits is *correlated* with the money spent on vegetables, but that there is actually a causal connection between them that exists only when the conservation requirement is in place. D-separation upon conditioning on a common causal child does not imply that any new causal connections form between the causal parents. Perhaps the best causal representation is to consider that the causal graph ‘Money spent on fruits→Total money spent←Money spent on vegetables’ is actually replaced by the causal graph ‘Money spent on fruits←Total money spent→Money spent on vegetables’ upon conditioning on ‘Total money spent’.

Systems that contain imposed conservation laws (conservation of energy, mass, volume, number, etc.) cannot yet be properly expressed using directed graphs and d-separation. In fact, such ‘causal’ relationships resemble Plato's notion of ‘formal causes’ rather that the ‘efficient causes’ with which scientists are used to working. However, it is important to keep in mind that this does not apply to conservation relationships that are due to a dynamic equilibrium, for which cyclic graphs can be used, but, rather, to conservation relationships that are imposed independently of the causal parents of the conserved variable.

2.13 Counter-intuitive consequences and limitations of d-separation: unfaithfulness

Let's go back to the relationship between d-separation and probabilistic independence. We now know that, once we have specified the acyclic causal model, every d-separation relation that exists in our causal model must be mirrored in an equivalent statistical independency in the observational data if the causal model is correct. This does not depend on any distributional assumptions of the random variables or on the functional form of the causal relationships. Is the contrary also true? If we find a (conditional) independency in our data then does this mean that there must always be a d-separation relationship in the causal process generating the data? The answer is: almost, but not, always. It is possible, as a limiting case, for there to be independencies in the data that are not predicted by the d-separation criterion. For instance, this can occur if the quantitative causal effect of two variables along different directed paths exactly cancel each other out. Two examples are shown in Figure 2.13. In these causal models we see that no vertex is unconditionally d-separated from any other vertex. Assume that the joint probability distribution over the three vertices is multivariate normal and that the functional relationships between the variables are linear. Under these conditions, we can use Pearson's partial correlation to measure probabilistic independence.[18] By definition, the partial correlation between X and Z, conditioned on Y, is given by:

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XY}\rho_{ZY}}{\sqrt{(1-\rho_{XY}^2)(1-\rho_{ZY}^2)}}$$

It can happen that $\rho_{XZ.Y} = 0$ (i.e. $\rho_{XZ} = \rho_{XY}\rho_{ZY}$) even though X and Z are not d-separated given Y, if the correlations between each pair of variables exactly cancel each other. Using the rules of path analysis (Chapter 4), this will happen only if Y is perfectly correlated with X in the first model in Figure 2.13, or if the indirect effect of X on Z is exactly equal in strength but opposite in sign to the direct effect of X on Z.

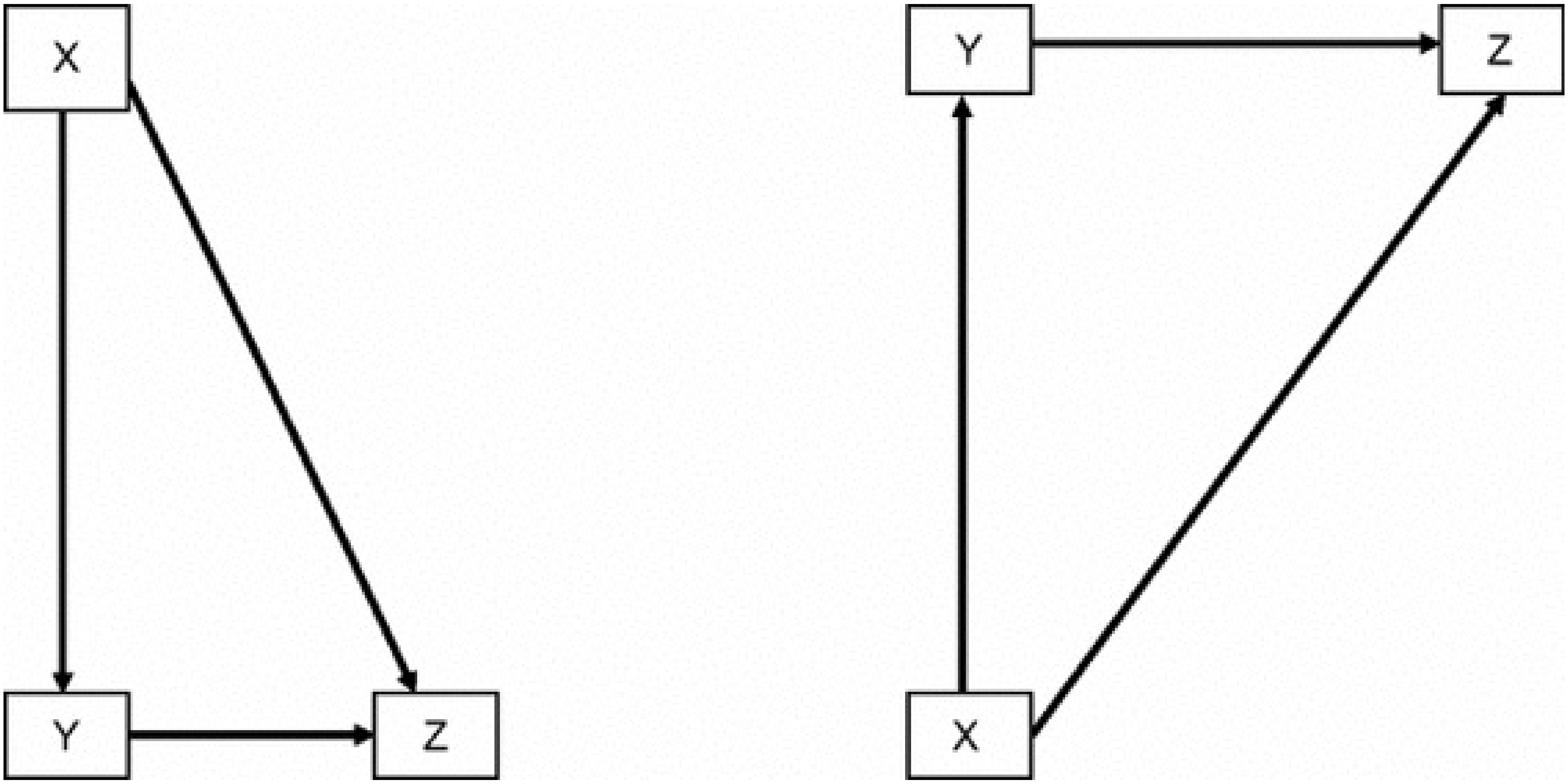


Figure 2.13 Two causal graphs for which special combinations of causal strengths can result in unfaithful probability distributions

When this occurs we say that the probability distribution is *unfaithful* to the causal graph (Pearl 1988; Spirtes, Glymour and Scheines 1993). I will call such probabilistic independencies that are not predicted by d-separation, and that depend on a particular combination of quantitative effects, *balancing independencies*, to emphasise that such independencies require a very peculiar balancing of the positive and negative effects between the variables along different paths. Clearly, this can occur only under very special conditions, and anyone who wanted to link a causal model with such an unfaithful probability distribution would require strong external evidence to support such a delicate balance of causal effects. This is not to say that such things are impossible. It sometimes happens that an organism attempts to maintain some constant set-point value by balancing different causal effects; an example is the control of the internal CO_2 concentration of a leaf, as described in Chapter 3. Essentially, in proposing such a claim we are saying that nature is conspiring to give the impression of independence by exactly balancing the positive and negative effects.

2.14 Counter-intuitive consequences and limitations of d-separation: context-sensitive independence

Another way in which independencies can occur in the joint probability distribution without being mirrored in the d-separation criterion is as a result of *context-sensitive* independence. An example of this in biology is enzyme induction.[19] Imagine a case in which the number of functional copies of a gene (G) determines the rate at which some enzyme is produced (E). If there are no functional copies of the gene then the enzyme is never produced. However, the rate at which these genes are transcribed is determined by the amount of some environmental inducer (I). If the environment completely lacks the inducer then no genes are transcribed and the enzyme is still never produced. It is possible to arrange an experimental set-up in which the number of functional genes is causally independent of the concentration of the inducer in the environment.[20] The number of functional genes and the concentration of the inducer are both causes of enzyme production. We can construct a causal graph of this process (Figure 2.14).

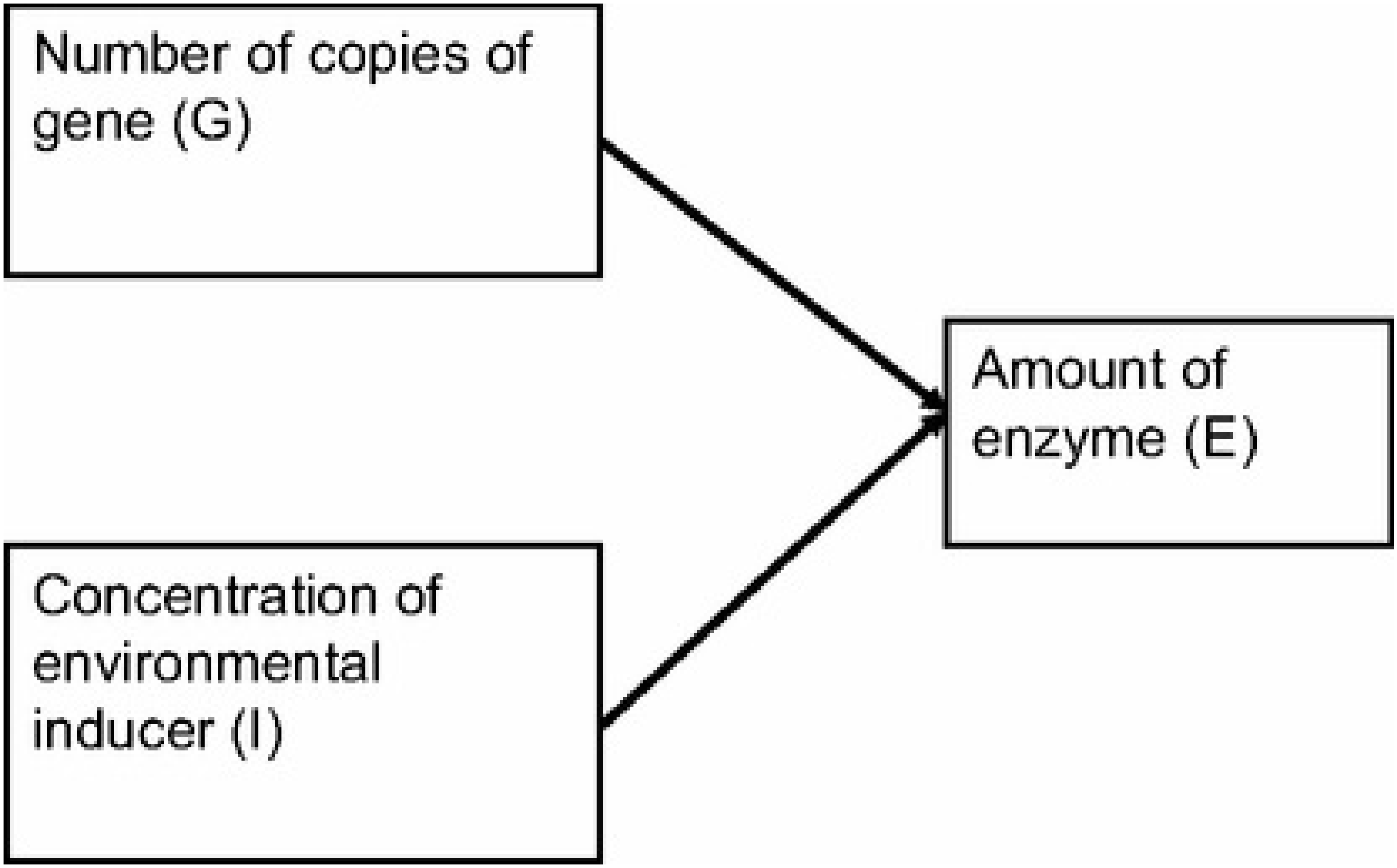

Number of copies of gene (G)
Amount of enzyme (E)
Concentration of environmental inducer (I)

Figure 2.14 A biological example of a causal process that can potentially result in context-sensitive independence

Now, applying d-separation to the causal graph in Figure 2.14 predicts that G is independent of I but that E is dependent on both G and I. However, if there are no copies of G (i.e. G = 0) then the concentration of the inducer will be independent of the amount of enzyme that is produced (which will be zero). Similarly, if there is no inducer (i.e. I = 0) then the number of copies of the gene will be independent of the amount of enzyme that is produced (which will be zero). In other words, for the special cases of G = 0 and/or I = 0, d-separation predicts a dependence when, in fact, there is independence. Note that the d-separation theorem still holds; d-separation does not predict any *independence* relations that do not exist. So long as the experiment involves experimental units, at least some of which include G ≠ 0 and I ≠ 0, the d-separation criterion still predicts both probabilistic independence and dependence. Similarly, if both G and I are true random variables (i.e. in which the experimenter has not fixed their values) then any reasonably large random sample will include such cases.

2.15 The logic of causal inference

Now that we have our translation device and are aware of some of the counter-intuitive results and limitations that can occur with d-separation, we have to be able to infer causal consequences from observational data by using this translation device. The details of how to carry out such inferences will occupy the rest of this book. However, before we look at the statistical details we must first consider the logic of causal and statistical inferences.

Since we are talking about the logic of inferences from empirical experience, it is useful to briefly look at what philosophers of science have had to say about valid inference. Logical positivism, itself rooted in the British empiricism of the nineteenth century that so influenced people such as Karl Pearson,[21] was dominant in the twentieth century up to the middle of the 1960s. This philosophical school was based on the verifiability theory of meaning; to be meaningful, a statement had to be of a kind that could be shown to be either true or false. For logical positivism, there were only two kinds of meaningful statements. The first kind was composed of *analytical* statements (tautologies, mathematical or logical statements) whose truth could be determined by deducing them from axioms or definitions. The second kind was composed of *empirical* statements that were either self-evident observations ('The water is 23°C') or could be logically deduced from combinations of basic observations whose truth was self-evident.[22] Thus, logical positivists emphasised the hypothetico-deductive method: a hypothesis was formulated to explain some phenomenon by showing that it followed deductively from the hypothesis. The scientist attempted to validate the hypothesis by deducing logical consequences of the hypothesis that were not involved in its formulation and testing these against additional observations. A simplified version of the argument goes like this:

- if my hypothesis is true then consequence C must also be true;
- consequence C is true;
- therefore, my hypothesis is true.

Readers will immediately recognise that such an argument commits the logical fallacy of affirming the consequent. It is possible for the consequence to be true even though the hypothesis that deduced it is false, since there can always be other reasons for the truth of C.

Popper (1980) has pointed out that, although we cannot use such an argument to verify hypotheses, we can use it to reject them without committing any logical fallacy:

- if my hypothesis is true then consequence C must also be true;
- consequence C is false;
- therefore, my hypothesis is false.

Practising scientists would quickly recognise that this argument, though logically acceptable, has important shortcomings when applied to empirical studies. It was recognised as long ago as the turn of the twentieth century (Duhem 1914) that no hypothesis is tested in isolation. Every time that we draw a conclusion from some empirical observation we rely on a whole set of auxiliary hypotheses (A_1, A_2, etc.) as well. Some of these have been repeatedly tested so many times and in so many situations that we scarcely doubt their truth. Other auxiliary assumptions may be less well established. These auxiliary assumptions will typically include ones concerning the experimental or observational background, the statistical properties of the data, and so on. Did the experimental control really prevent the variable from changing? Were the data really normally distributed, as the statistical test assumes? Such auxiliary assumptions are legion in every empirical study, including the randomised experiment, the controlled experiment or the methods described in this book involving statistical controls. A large part of every empirical investigation involves checking, as best as one can, such auxiliary assumptions so that, once the result is obtained, blame or praise can be directed at the main hypothesis rather than at the auxiliary assumptions.

Popper's process of inference might therefore be simplistically paraphrased[23] as:

- if auxiliary hypotheses A_1, A_2,…A_n are true, and
- if my hypothesis is true then consequence C must be true;
- consequence C is false;
- therefore, my hypothesis is false.

Unfortunately, to argue in such a manner is also logically fallacious. Consequence C might be false not because the hypothesis is false but, rather, because one or more of the auxiliary hypotheses are false. The empirical researcher is now back where he or she started: there is no way of determining either the truth or falsity of his or her hypothesis in any absolute sense from logical deduction. This conclusion applies just as well to the randomised experiment, the controlled experiment or the methods described in this book. However, most biologists would recognise the falsifiability criterion as important to science, and would probably modify the simplistic paraphrase of Popper's inference by attempting to judge which – the auxiliary hypotheses and background conditions or the hypothesis under scrutiny – is on firmer empirical ground. If the auxiliary assumptions seem more likely to be true than the hypothesis under scrutiny, yet if the data do not accord with the predicted consequences then the hypothesis would be tentatively rejected. If there are no reasoned arguments to suggest that the auxiliary assumptions are false, and the data also accord with the predictions of the hypothesis under scrutiny, then the hypothesis would be tentatively accepted.

Pollack (1986) calls such reasoning *defeasible* reasoning.[24] Revealingly, practising scientists have explicitly described their inferences in such terms for a long time. At the turn of the twentieth century Thomas Huxley likened the decision to accept or reject a scientific hypothesis to a criminal trial in a court of law (reproduced by Rapport and Wright 1963) in which guilt must be demonstrated beyond reasonable doubt.

Let's apply this reasoning to the examples in Chapter 1 involving the randomised and the controlled experiments. Later, I will apply the same reasoning to the methods involving statistical control.

Here is the logic of causal inference with respect to the randomised experiment to test the hypothesis that fertiliser addition increases seed yield.

- If the randomisation procedure was properly done so that the alternate causal explanations were excluded;
- if the experimental treatment was properly applied;
- if the observational data do not violate the assumptions of the statistical test;
- if the observed degree of association was not due to sampling fluctuations;
- then by the causal hypothesis the amount of seed produced will be associated with the presence of the fertiliser.
- There is/is not an association between the two variables.
- Therefore, the fertiliser addition might have caused/did not cause the increased seed yield.

This list of auxiliary assumptions is only partial. In particular, we still have to make the basic assumption linking causality to observational associations, as described in Chapter 1. At this stage we must either reject one of the auxiliary assumptions or tentatively accept the conclusion concerning the causal hypothesis. If the probability associated with the test for the association is sufficiently large, traditionally above 0.05,[25] then we are willing to reject one of the auxiliary assumptions (the observed measure of association was not due to sampling fluctuations) rather than accept the causal hypothesis. Thus, we reject our causal hypothesis. This rejection must remain tentative. This is because another of the auxiliary assumptions (not listed above) is that the sample size is large enough to permit the statistical test to differentiate between sampling fluctuations and systematic differences. However, note that it is not enough to propose any old reason to reject one of the auxiliary assumptions; we must propose a reason that has empirical support. We must produce *reasonable* doubt; in the context of the assumption concerning sampling fluctuations, scientists generally require a probability above 0.05.

Here it is useful to quote from the first edition of Fisher's (1925: 504) influential *Statistical Methods for Research Workers*: ‘Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.’ It is clear that Fisher was demanding reasonable doubt concerning the null hypothesis, since he asked only that a result ‘rarely fail’ to reject it. What if the probability of the statistical test was sufficiently small, such as 0.01, that we do not have reasonable grounds to reject our auxiliary assumption concerning sampling fluctuations? What if we do not have reasonable grounds to reject the other auxiliary assumptions? What if the sampling variation was small compared to a reasonable effect size? Then we must tentatively accept the causal hypothesis.

Again, this acceptance must remain tentative, since new empirical data might provide such reasonable doubt. Is there any automatic way of measuring the relative support for or against each of the auxiliary assumptions and of the principal causal hypothesis? No. Although the support (in terms of objective probabilities) for some assumptions can be obtained – for instance, those concerning the normality or linearity of the data – there are many other assumptions that deal with experimental procedure or a lack of confounding variables for which no objective probability can be calculated. This is one reason why so many contemporary philosophers of science prefer Bayesian methods to frequency-based interpretations of probabilistic inference (see, for example, Howson and Urbach 1989). Such Bayesian methods suffer from their own set of conceptual problems (Mayo 1996). In the end, even the randomised experiment requires subjective decisions on the part of the researcher. This is why the independent replication of experiments in different locations, using slightly different environmental or experimental conditions and therefore having different sets of auxiliary assumptions, is so important. As the causal hypothesis continues to be accepted in these new experiments, it becomes less and less reasonable to suppose that incorrect auxiliary assumptions are conspiring to give the illusion of a correct causal hypothesis.

Here is the logic of our inferences with respect to the controlled experiment to test the hypothesis that renal activity causes the change in the colour of the venous renal blood, described in Chapter 1.

- If the activity of the kidney was effectively controlled;
- if the colour of the blood was accurately determined;
- if the experimental manipulation did not change some other uncontrolled attribute besides kidney function that is a common cause of the colour of blood in the renal vein before entering, and after leaving, the kidney;
- if there was not some unknown (and therefore uncontrolled) common cause of the colour of blood in the renal vein before entering, and after leaving, the kidney;
- if a rare random event did not occur;
- then, by the causal hypothesis, blood will change colour only when the kidney is active.
- The blood did change colour in relation to kidney activity.
- Therefore, kidney activity does cause the change in the colour of blood leaving the renal vein.

Again, this list of auxiliary assumptions is only partial. Again, one must either produce reasonable evidence that one or more of the auxiliary assumptions is false or tentatively accept the hypothesis. In particular, more of these auxiliary assumptions concern properties of the experiment or of the experimental units for which we cannot calculate any objective probability concerning their veracity. This was one of the primary reasons why Fisher rejected the controlled experiment as inferior. In the controlled experiment these auxiliary assumptions are more substantial, but it is still not enough to raise any old doubt; there must be some empirical evidence to support the decision to reject one of these assumptions. Since we want the data to cast doubt or praise on the principal causal hypothesis and not on the auxiliary assumptions, we will ask only for evidence that casts reasonable doubt. It is not enough to reject the causal hypothesis simply because ‘experimental manipulation *might have* changed some other uncontrolled attribute besides kidney function that is a common cause of the colour of blood in the renal vein before entering, and after leaving, the kidney’. We must advance *some* evidence to support the idea that such an uncontrolled factor in fact exists. For instance, a critic might reasonably point out that some other attribute is also known to be correlated with blood colour and that the experimental manipulation was known to have changed this attribute.

Although such evidence would certainly not be sufficient to demonstrate that this other attribute definitely was the cause, it might be enough to cast doubt on the veracity of the principal hypothesis.

This is the same criterion that we used beforehand to choose a significance level in our statistical test. Rejecting a statistical hypothesis because the probability associated with it was, say, 0.5 would not be reasonable. Certainly, this gives some doubt about the truth of the hypothesis, but our doubt is not sufficiently strong that we would have a clear preference for the contrary hypothesis. It is the same defeasible argument that might be raised in a murder trial. If the prosecution has demonstrated that the accused had a strong motive, if it produced a number of reliable eyewitnesses and if it produced physical evidence implicating the accused then it would not be enough for the defence to simply claim that ‘maybe someone else did it’. However, if the defence could produce some contrary empirical evidence implicating someone else then reasonable doubt would be cast on the prosecution's argument. In fact, I think that the analogy between testing a scientific hypothesis and testing the innocence of the accused in a criminal trial can be stretched even further. There is no objective definition of reasonable doubt in a criminal trial; what is reasonable is decided by the jury in the context of legal precedence. In the same way, there is no objective definition of reasonable doubt in a scientific claim. In the first instance, reasonable doubt is decided by the peer reviewers of the scientific article, and, ultimately, reasonable doubt is decided by the entire scientific community. One should not conclude from this that such decisions are purely subjective acts and that scientific claims are therefore simply relativistic stories whose truth is decided by fiat by a power elite. Judgements concerning reasonable doubt and statistical significance are constrained in that they must deliver predictive agreement with the natural world in the long run.

Now let's look at the process of inference with respect to causal graphs.

- If the data were generated according to the causal model;
- if the causal process generating the data does not include non-linear feedback relationships;
- if the statistical test used to test the independence relationships is appropriate for the data;
- if a rare sampling fluctuation did not occur;
- then each d-separation statement will be mirrored by a probabilistic independence in the data.
- At least one predicted probabilistic independence did not exist.
- Therefore, the causal model is wrong.

By now, you should have recognised the similarity of these inferences. We can prove by logical deduction that d-separation implies probabilistic independence in such directed acyclic graphs. We can prove that, barring the case of non-linear feedback with non-normal data (an auxiliary assumption), every d-separation statement obtained from any DAG must be mirrored by a probabilistic independence in any data that were generated according to the causal process that was coded by this DAG. We can prove that, barring a non-faithful probability distribution (another auxiliary assumption, but one that is relevant only if the causal hypothesis is accepted, not if it is rejected), there can be no independence relation in the data that is not mirrored by d-separation. So, if we have used a statistical test that is appropriate for our data and have obtained a probability that is sufficiently low to reasonably exclude a rare sampling event, we must tentatively reject our causal model. As in the case of the controlled experiment, if we are led to tentatively accept our causal model then this will require that we can't reasonably propose an alternative causal explanation that also fits our data as well. As always, it is not sufficient to simply claim that '*maybe* there is such an alternative causal explanation'. One must be able to propose an alternative causal explanation that has at least enough empirical support to cast reasonable doubt on the proposed explanation.

2.16 Statistical control is not always the same as physical control

We have now seen how to translate from a causal hypothesis into a statistical hypothesis. First, transcribe the causal hypothesis into a causal graph showing how each variable is causally linked to other variables in the form of direct and indirect effects. Second, use the d-separation criterion to predict what types of probabilistic independence relationships must exist when we observe a random sample of units that obey such a causal process. In Chapter 1 I alluded to the fact that the key to a controlled experiment is *control* over variables, not how the control is produced. It is time to look at this more carefully. The relationship between control through external (experimental) manipulation and probability distributions is given by the manipulation theorem (Spirtes, Glymour and Scheines 1993). Let me introduce another definition.

> **Back-door path**: given two variables, X and Y, and a variable F that is a causal ancestor of both X and Y, a back-door path goes from F to each of X and Y. Thus X← ← ←F→ → →Y.

Whenever someone directly physically controls some set of variables through experimental manipulation, he or she is changing the causal process that is generating the data. Whenever someone physically fixes some variable at a given level the variable stops being random[26] and is then under the complete control of the experimenter. In other words, whatever causes might have determined the 'random' values of the variable before the manipulation have been removed by the manipulation. The only direct cause of the controlled variable after the manipulation has been performed is the will of the experimenter.

Imagine that someone has randomly sampled herbaceous plants growing in the understorey of an open stand of trees. The measured variables are the light intensities experienced by the herbaceous plants, their photosynthetic rates and the concentration of anthocyanins (red-coloured pigments) in their leaves. Each of these three variables is random since they are outside the control of the researcher. One cause of variation in light intensity at ground level is the presence of trees. The researcher proposes two alternative causal explanations for the data (Figure 2.15).

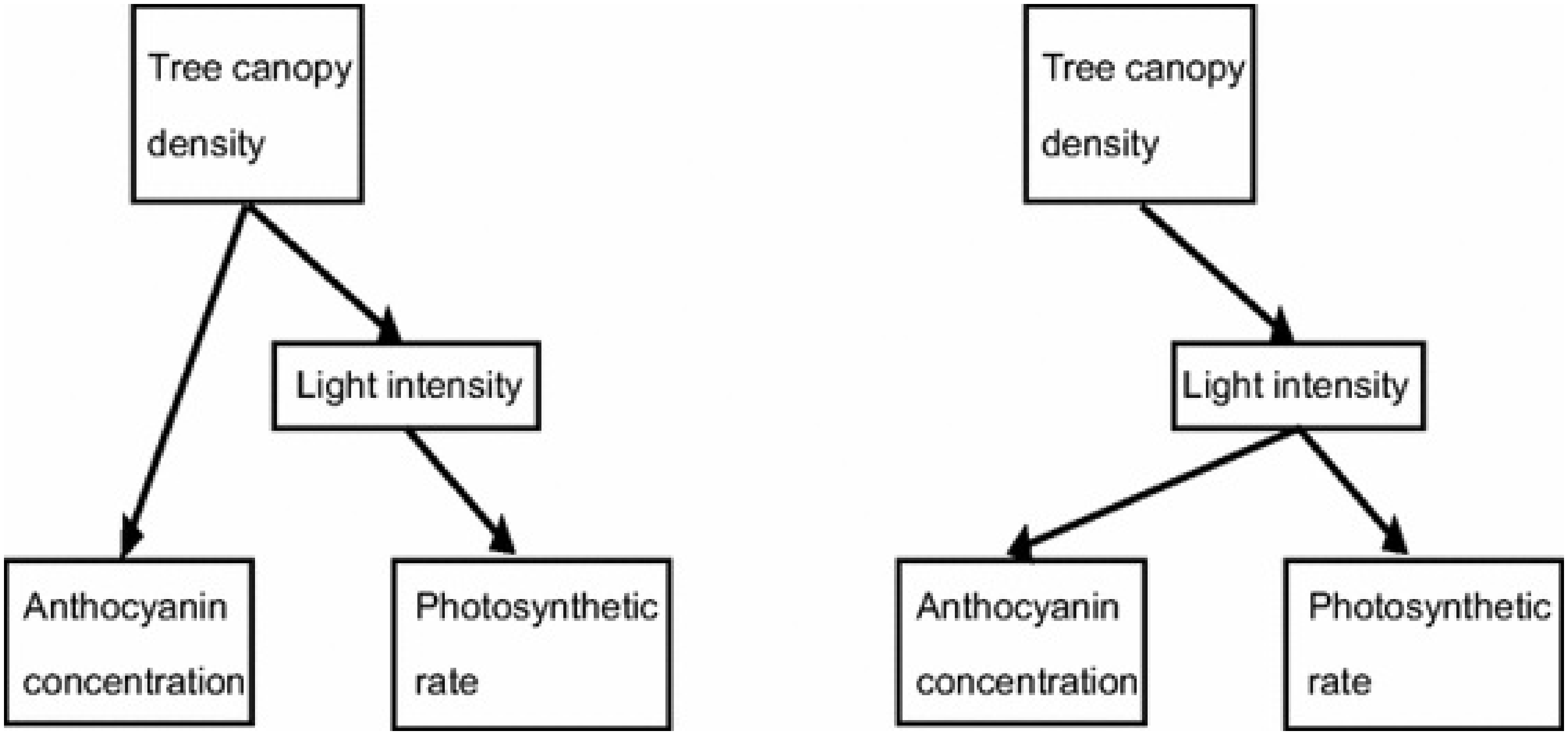


Figure 2.15 Two different causal scenarios linking the same four variables

To test between these two alternative explanations, the researcher experimentally manipulates light intensity by installing a neutral shade cloth between the trees and the herbs, and then adds an artificial source of lighting. Remembering that this is a controlled experiment, the researcher would want to take precautions to ensure that other environmental variables (temperature, humidity, etc.) are not changed by this manipulation. The manipulation theorem, in graphical terms,[27] states that the probability distribution of this new causal system can be described by taking the original (unmanipulated) causal graphs, removing any arrows leading into the manipulated variable (light intensity) and adding a new variable representing the new causes of the manipulated variable (Figure 2.16).

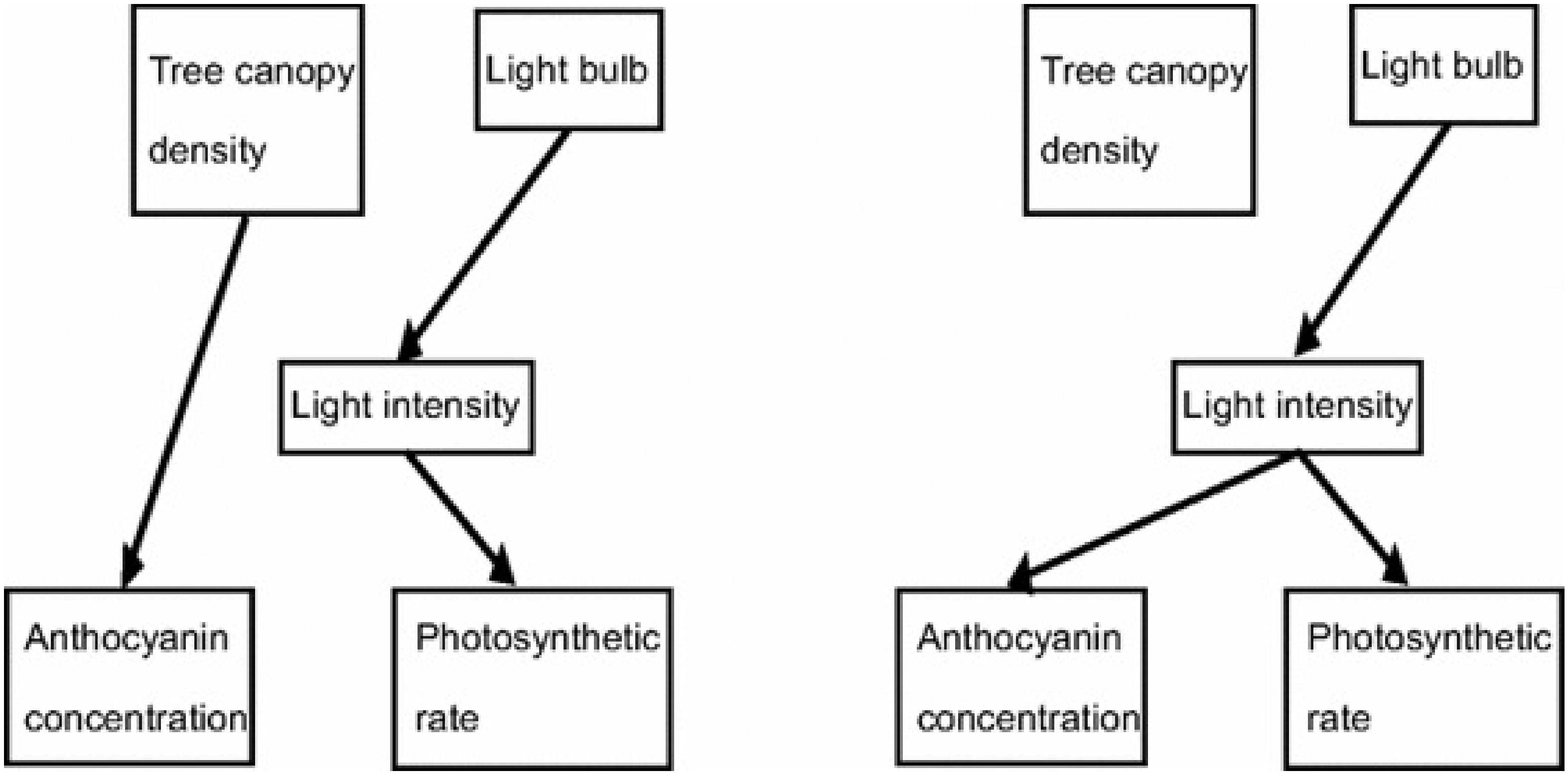


Figure 2.16 Experimental manipulation of the causal systems that are shown in Figure 2.15

D-separation will predict the pattern of probabilistic independencies in this new causal system. Notice that anthocyanin concentration is d-separated from the photosynthetic rate according to the first hypothesis in both the manipulated system (Figure 2.16), when light intensity is experimentally fixed, and the unmanipulated system (Figure 2.15), when light intensity is statistically fixed by conditioning. The same d-connection relationships between anthocyanin concentration and the photosynthetic rate hold in the second scenario whether based on physically or statistically controlling light intensity. In other words, statistical and experimental controls are alternative ways of doing the same thing: predicting how the associations between variables will change once other sets of variables are 'held constant'. This does not mean that the two types of control always predict the same types of observational independencies in our data; remember the example of d-separation upon conditioning on a causal child, described previously. Once we have a way of measuring how closely the predictions agree with the observations, we have a way of testing, and potentially falsifying, causal hypotheses even in cases in which we cannot physically control the variables of interest.

With these notions we can now go back and look again at the randomised experiment in Chapter 1. Let's consider an example involving an agricultural researcher who is interested in determining if, and how, the addition of a nitrate fertiliser can increase plant yield. To be more specific, imagine that the plant is alfalfa, which contains a bacterium in its roots that is capable of directly fixing atmospheric nitrogen (N_2). The researcher meets a farmer, who tells him that adding such a nitrate fertiliser in the past had increased the yield of alfalfa. After further questioning, the researcher learns that the farmer had tended to add more fertiliser to those parts of the field that, in previous years, had produced the lowest yields. The researcher knows that other things can also affect the amount of fertiliser that a farmer will add to different parts of a field. For instance, parts of the field that cause the farmer to slow down the speed of his tractor will therefore tend to receive more fertiliser, and so on.[28] Imagine that, unknown to the researcher, the actual causal processes are as shown in Figure 2.17. There are only three sources of nitrogen: the nitrate that is added to the soil by the fertiliser, by NO_X deposition and from N_2 fixation by the bacterium. The amount of fertiliser added by the farmer in different parts of the field is determined by the yield of plants the previous year as well as the contours of the field. In reality, all the sources of nitrogen and the soil phosphate level are causes of yield.

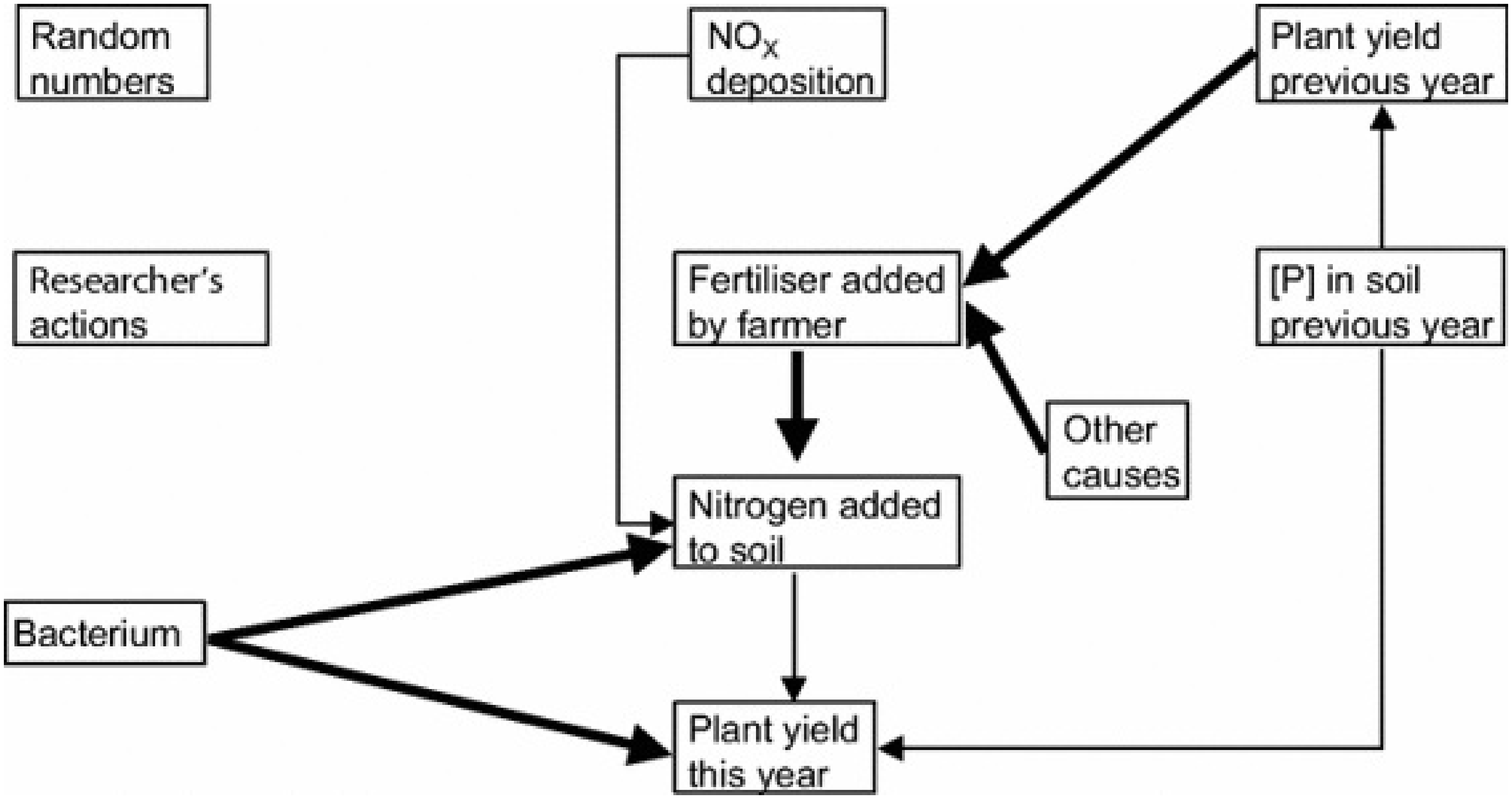


Figure 2.17 A hypothetical causal system before experimental manipulation

Before experimenting with this system, the researcher has previous causal knowledge of only part of it, shown by the thicker arrows in Figure 2.17. He knows that the bacterium will increase the alfalfa yield. He knows that the bacterium will increase the nitrate concentration in the soil. He knows that the yield of alfalfa in previous years has affected the amount of nitrate fertiliser that the farmer had added, and he knows that the amount of added nitrate fertiliser is *associated* with increased yields. What he doesn't know is whether the nitrogen added to the soil is the cause of the subsequent plant yield.

Since the experiment has not yet begun, the ‘Random numbers’ in Figure 2.17 do not affect any actions by the researcher, and he has no causal effect on any variable in the system. The ‘Random numbers’ and the ‘Researcher's actions’ are therefore causally independent of each other and of every other variable in the system.

Based only on the *partial* knowledge shown by the thick arrows, can the researcher use d-separation and statistical control to confidently infer that the added nitrate fertiliser causes an increase in plant yield? No. He knows that the yields of previous years were a cause of the farmer's fertiliser addition and not vice versa; therefore, he knows that he can block any possible back-door path between the amount of fertiliser added and the plant yield that passes through the variable ‘Plant yield previous year’. Unfortunately, he also knows that this was not the *only* possible cause of the amount of fertiliser added by the farmer to different parts of the field. Therefore, he cannot exclude the possibility that there is some back-door path that does not include the variable ‘Plant yield previous year’ and that is generating the association between the present plant yield and the amount of fertiliser added by the farmer. Remember that, to invoke such a possibility, one must be able to present some empirical evidence that such a back-door path might exist, but this would be easy to do. For instance, if the tractor slows down[29] as it begins to go up a slope (and therefore deposits more fertiliser), and if water (which is known to increase plant yield) tends to accumulate at the bottom of the slope, then we have a possible back-door path (fertiliser added←tractor slowed down←hill→water accumulation→plant yield).

The researcher knows that it is possible to randomly assign different levels of nitrate fertiliser to plots of ground in a way that is not caused by any attribute of these plots. He persuades the farmer not to add any fertiliser. The previous cause of the amount of fertiliser added has been erased in this new context, and so the arrow from ‘Plant yield previous year’ to ‘Fertiliser added by farmer’ is removed from the causal graph. Since the farmer has agreed not to add any fertiliser, the value of this variable is fixed at zero, and so all arrows coming out of this variable are also erased. The researcher decides to add nitrate fertiliser to different plots at either 0 or 20 kg/hectare based only on the value of randomly chosen numbers. Therefore, we add an arrow from ‘Random numbers’ to ‘Researcher's actions’ and an arrow from ‘Researcher's actions’ to ‘Nitrogen added to soil’. Remember that an arrow signifies a *direct* cause – i.e. a causal effect that is not mediated through other variables in the causal explanation. Consequently, we cannot add an arrow from ‘Researcher's actions’ to ‘Plant yield this year’ unless we believe that the researcher's actions do cause a change in plant yield this year and that this cause is not completely mediated by some other set of variables in the causal system. The causal structure that exists after the experimental manipulation is shown in Figure 2.18.

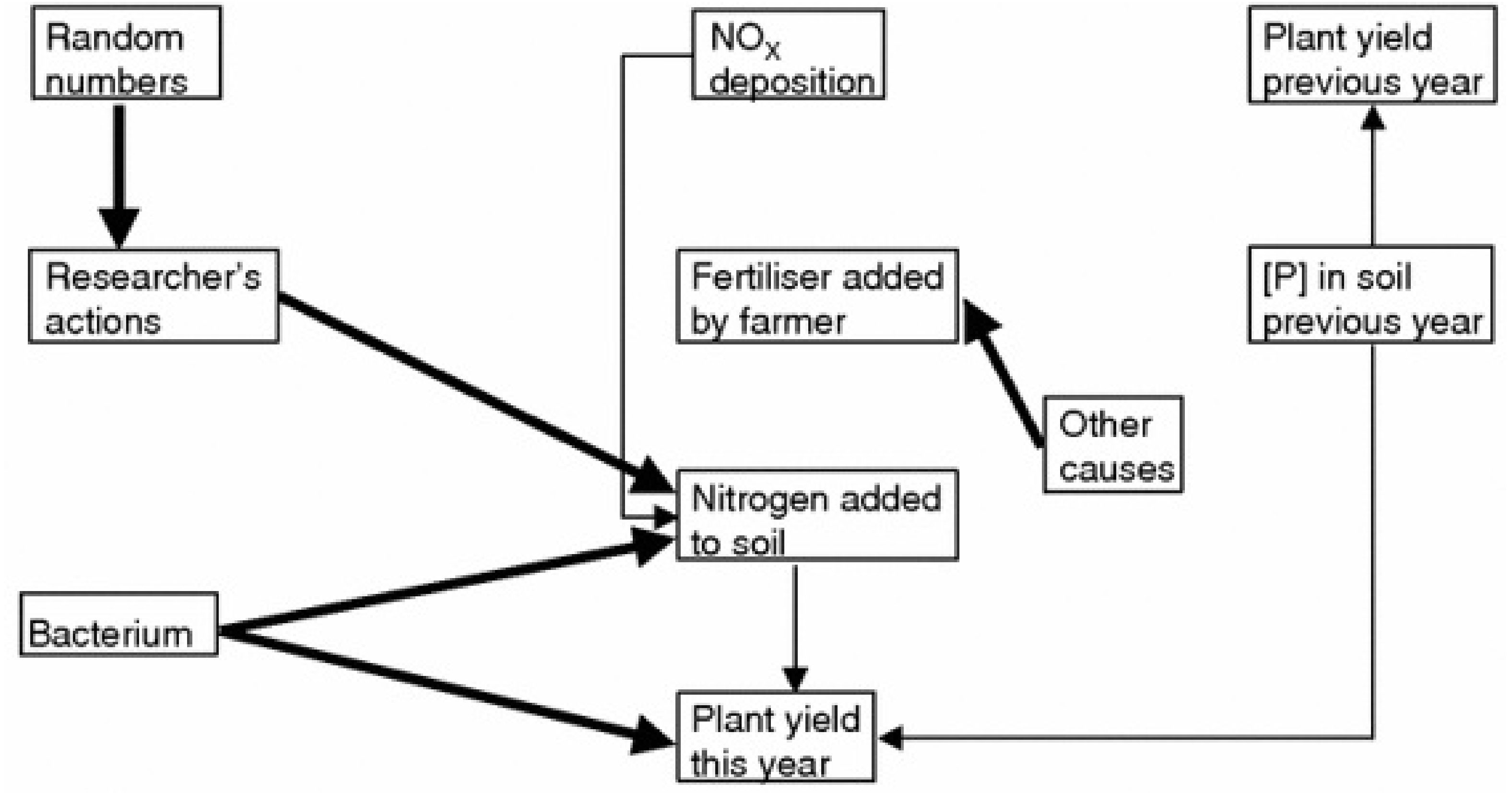


Figure 2.18 Experimental manipulation of the causal system shown in Figure 2.17 based on a randomised experiment

Given this new causal scenario, we can now use d-separation to determine if there is a causal relationship between the amount of nitrate fertiliser added by the researcher and the plant yield that year. If one can trace a directed path beginning at ‘Researcher's actions’ and passing through ‘Plant yield this year’ by following the direction of the arrows then the two are not d-separated. This necessarily implies that there will be a statistical association between the two variables. If no such directed path exists then the addition of nitrate fertiliser by the researcher does not cause a change in plant yield this year. In fact, these two variables are not d-separated in this causal graph, and so such a randomised experiment would detect an effect of fertiliser addition on plant yield. In Chapter 1 I said that if there is a statistical association between two variables, X and Y, then there can be only three elementary (but not mutually exclusive) causal explanations: either X causes Y (shown by a directed path leading from X and passing into Y), Y causes X (shown by a directed path leading from Y and passing into X) or there is some other variable (F) that is a cause of both X and Y (shown by a back-door path from F and into both X and Y). Because the researcher has agreed to act completely in accordance with the results of the randomisation process, we know that no arrows point into ‘Researcher's actions’ except the one coming from ‘Random numbers’. The random numbers are not caused by any attribute of the system. Therefore, the researcher knows that there can be no back-door paths confounding his results because he knows that there are no arrows pointing into ‘Researcher's actions’ except for the one coming from ‘Random numbers’. If there is a statistical association between ‘Researcher's actions’ and ‘Plant yield this year’ that cannot reasonably be attributed to random sampling fluctuations then the researcher knows that the association must be due to a directed path coming from ‘Researcher's actions’ and passing through ‘Plant yield this year’. This is why such a randomised experiment, in conjunction with a way of calculating the probability of observing such a random event, can provide a strong inference concerning a causal effect. The reader should note that even the randomisation process might not allow the researcher to conclude that ‘Nitrogen added to the soil’ is a *direct* cause of increased plant yield. In Figure 2.18 the researcher has already concluded that there is a back-door path from these two variables emanating from the presence of the nitrogen-fixing bacterium, and so to make such a claim he would have to provide evidence beyond a reasonable doubt that his actions did not somehow affect the abundance or activity of these bacteria.

Now, let's modify the causal scenario a bit. Imagine that the farmer has agreed to let the researcher conduct an experiment and promises not to add any fertiliser while the experiment is in progress, but insists that the researcher ensure that the parts of the field that had produced the lowest plant yield last year must absolutely receive more fertiliser this year. The researcher decides to allocate the fertiliser treatment in the following way: after choosing the random numbers as before, he also adds 5 kg/hectare to those plots whose previous yields were below the median value. Figure 2.19 shows this causal scenario. By doing so he is no longer conducting a true randomised experiment.

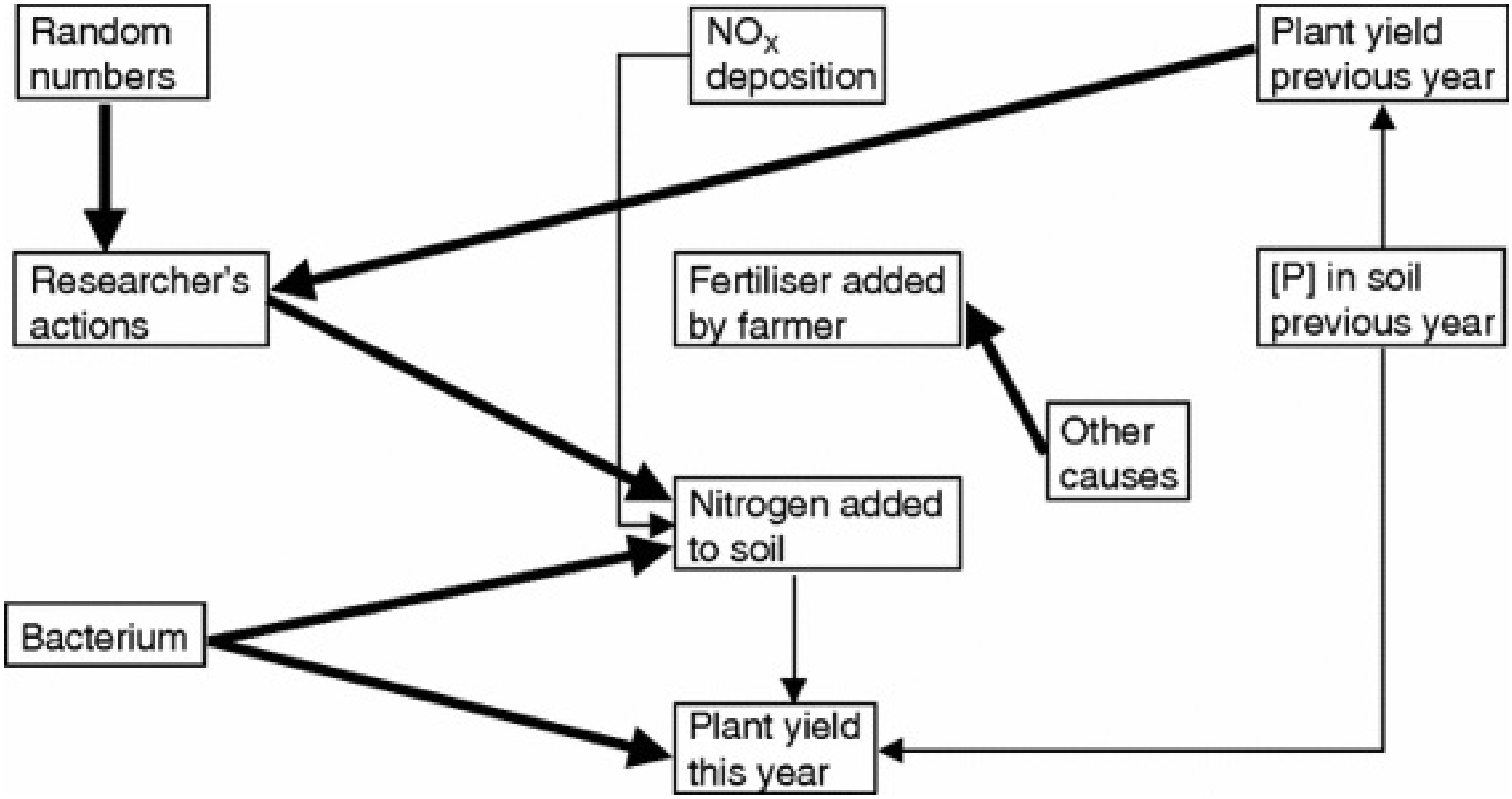


Figure 2.19 Experimental manipulation of the causal system shown in Figure 2.17 that is not based on a randomised experiment

Now, using d-separation we see that there would be an association between 'Researcher's actions' and 'Plant yield this year' even if there was no causal effect of the amount of nitrate fertiliser added and the plant yield that follows. The reason is that there is now a back-door path linking the two variables through the common cause '[P] in soil previous year'. This path has been created by allowing 'Plant yield previous year' to be a cause of the researcher's actions. Yet all is not lost. He systematically assigned fertiliser levels based *only* on the yield data of the previous year plus the random numbers. This means that he knows that there are only two independent causes determining how much fertiliser each plot received. He also knows, because of d-separation, that any causal signal passing from any unknown variable into 'Researcher's actions' through 'Plant yield previous year' is blocked if he statistically controls for 'Plant yield previous year'. He can make this causal inference without knowing anything else about the causal system. Therefore, he knows that, once he statistically conditions on 'Plant yield previous year', any remaining statistical association, if it exists, must be due to a causal signal coming from 'Researcher's actions' and following a directed path into 'Plant yield this year'. This causal inference is just as solid as in the previous example, in which treatment allocation was due only to random numbers. What allows him to do this in this controlled, but not strictly randomised, experiment but not in the original non-manipulated system in which the farmer applied the fertiliser based on previous yield data? If you compare Figures 2.17 (non-manipulated) and 2.19 (controlled, non-randomised manipulation) you will see that in Figure 2.17 there were other causes, besides yield, that influenced the farmer's actions. These other causes were both unknown and unmeasured, thus preventing the researcher from statistically controlling for them, and this left open the possibility of other back-door paths that would confound the causal inference. In Figure 2.19 the experimental design ensured that the only cause (i.e. previous yields) was already known and measured.

Using either randomised experiments or this controlled approach, the researcher could conclude[30] that his action of adding nitrate fertiliser does cause a change in the alfalfa yield and in the amount of nitrate in the soil.

Under what conditions could he infer that the *soil* nitrate levels (as opposed to nitrate fertiliser *addition*) causes the change in the alfalfa yield? In other words, what would allow him to infer that the fertiliser addition increased soil nitrate concentration, which, in turn, increased the alfalfa yield? Although he was able to randomise and to exert experimental control over the amount of fertiliser added to the soil, this is not the same as randomly assigning values of soil nitrate to the plots, and he has not exerted *direct* experimental control over soil nitrate levels. Because of this he cannot unambiguously claim that the experiment has demonstrated that soil nitrate levels cause an increase in plant yield. In other words, there might be a back-door path from the fertiliser addition to each of soil nitrate and plant yield even though soil nitrate levels may have had no direct effect on plant yield. For instance, perhaps the fertiliser addition reduced the population level of some soil pathogen whose presence was reducing plant growth.

He can test the hypothesis that the association between soil nitrate levels and plant yield is due only to a back-door path emanating from the amount of added fertiliser by measuring soil nitrate levels and then statistically controlling for this variable. D-separation predicts that, if this new causal hypothesis is true, the effect of fertiliser addition will still exist. If the effect of fertiliser addition was due only to its effect on soil nitrate levels then d-separation predicts that the effect of fertiliser addition on plant yield will disappear once the soil nitrate level is statistically controlled. Since he knows, from previous biological knowledge, that there is at least one back-door path linking soil nitrate and plant yield (due to the effect of the nitrogen-fixing bacteria in the root nodules) then he can determine if there is some other common cause generating a back-door path if he can measure and then control for the amount of this bacterium.

2.17 A taste of things to come

Up to now we have been inferring the properties of the observational model (the joint probability distribution) given the causal model that generates it. Can we also do the contrary? If we know the entire pattern of statistical independencies and conditional independencies in our observational model, can we specify the causal structure that must have generated it? No. It is possible for different causal structures to generate the same set of d-separation statements and, therefore, the same pattern of independencies. Nonetheless, it is possible to specify a *set* of causal models that all predict the same pattern of independencies that we find in the probability distribution; these are called *equivalent* models, and these are described in Chapter 8. By extension, we can exclude a vast group of causal models that could not have generated the observational data. There are two important consequences of this.

First, after proposing a causal model and finding that our observational data are consistent with it (i.e. that the data do not contradict any of the d-separation statements of our causal model), we can determine which other causal models would also be consistent with our data.[31] By definition, our data cannot distinguish between such equivalent causal models, and so we will have to devise other sorts of observations to differentiate between them.

Second, we can exploit the independencies in our observational data to generate such equivalent models even if we do not yet have a causal model that is consistent with our data. This leads to the topic of exploratory methods, which is discussed in Chapter 8. Such exploratory methods are very useful when theory is not sufficiently well developed to allow us to propose a causal explanation – a condition that occurs often in organismal biology.

However, before delving into these topics, we must first look at the mechanics of fitting such observational models, generating their correlational ‘shadows’ and comparing the observed shadows (the patterns of correlation and partial correlation) to the predicted shadows. This leads into the topic of path models and, more generally, structural equations. Chapters 3 to 7 deal with these topics.

[1] Fisherian statistics does deal with causal hypotheses, but the causal inferences come from the experimental design, not from the mathematical details; see Chapter 1.

[2] My children seem to have mastered this metaphysical concept well before age five. This is another example of how deeply ingrained the notion of causality is.

[3] This was written in the first edition. Current translation programs are much better!

[4] More accurately, directed graphs can economically store the conditional independence constraints implied by a causal system of an arbitrary joint probability distribution. This is explained in more detail below.

[5] In the jargon of graph theory, an undirected graph consists of a set of vertices {A,B,C,…} and a binary set denoting the presence or absence of edges (lines) between each pair of vertices. The graph becomes directed when we include a set of symbols for each edge showing direction. It is also possible to construct partially directed graphs. A graph is acyclic if there are no paths that lead a vertex back onto itself; otherwise it is cyclic. The causal graph in Figure 2.5 is therefore a directed acyclic graph (DAG).

[6] D-separation can also be extended to determining the causal independence of two sets of vertices **A** and **B**, upon conditioning on a third set **W**.

[7] In this case, a bivariate normal distribution.

[8] This can be generalised to joint distributions of sets of variables **X** and **Y** conditional on another set **Z**.

[9] $P(v_i)=\Pi P(v_i|\text{parents}(v_i))$.

[10] Fertiliser→photosynthetic enzymes→photosynthetic rate.

[11] This is explained later in this chapter.

[12] Such simulations are often called Monte Carlo simulations, after the famous gambling city, because they make use of random number generators to simulate a random process.

[13] Many commercial statistical packages can generate random numbers from specified probability distributions. A good reference, along with FORTRAN subroutines, is the book by Press et al. (1986). The R language incorporates most of these, of which the ‘rnorm’ function is one.

[14] Remember that a partial regression coefficient is a function of the partial correlation coefficient. The partial correlation coefficient measures the degree of linear association between two variables upon conditioning on some other set of variables; see Chapter 3.

[15] Even as a prediction device, such models are valid only if no manipulations are done to the population.

[16] On the other hand, if this process were to be repeated for a number of generations and the two attributes were heritable, there would develop a causal link, since the average values of the attributes in the next generation would depend on who survives, and this is a consequence of these same attributes in the previous generation.

[17] In the literature of structural equation modelling, cyclic or feedback models are called 'non-recursive'. This whole subject area is replete with confusing and intimidating jargon.

[18] Pearson partial correlations are explained more fully in Chapter 3.

[19] A classic example is the *lac* operon of *E. coli*, whose transcription in the presence of lactose induces the production of β-galactosidase, lac permease and transacetylase, thus converting lactose into galactose and glucose (De Robertis and De Robertis 1980).

[20] Whether this would be true in the biological population is an empirical question. Perhaps the presence of a functional gene was selected on the basis of the presence of the inducer. In this case, the inducer would be a cause of the presence (and perhaps the number of copies) of the gene.

[21] This is explored in more detail in Chapter 3.

[22] That even such simple observational or experiential statements cannot be considered objectively self-evident was shown at the beginning of the twentieth century by Duhem (1914).

[23] *Simplistic* because it is wrong. Popper did not make such a claim.

[24] *Defeasible* because it can be *defeated* with subsequent evidence.

[25] See Cowles and Davis (1982b) for a history of the 5 per cent significance level. The first edition of Fisher's classic book states (Fisher 1925: 47): 'It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.' The words 'convenient' and 'formal' emphasise the somewhat arbitrary nature of this value. In fact, this level can be traced back even further to the use of three times the probable error (about two-thirds of a standard deviation). Strictly speaking, twice the standard deviation of a normal distribution gives a probability level of 0.0456; perhaps Fisher simply rounded this up to 0.05 for his tables. Pearson and Kendall (1970) record Pearson's reasons at the turn of the century: $p = 0.5586$, 'thus we may

consider the fit remarkably good'; p = 0.28, 'fairly represented'; p = 0.10, 'not very improbable'; p = 0.01, 'this very improbable result'. Note that some doubt began at 0.1 and Pearson was quite convinced at p = 0.01. The midpoint between 0.1 and 0.01 is 0.05. Cowles and Davis (1982a) conducted a small psychological experiment by fooling students into believing that they were participating in a real betting game (with money) that was, in reality, fixed. The object was to see how unlikely a result people would accept before they began to doubt the fairness of the game. They found that, 'on average, people do have doubts about the operation of chance when the odds reach about 9 to 1 [i.e. 0.09], and are pretty well convinced when the odds are 99 to 1 [i.e. ~0.0101]…If these data are accepted, the 5% level would appear to have the appealing merit of having some grounding in common sense.'

[26] The notion of 'randomness' is another example of a concept that is regularly invoked in science even though it is extraordinarily difficult to define.

[27] The manipulation theorem also predicts how the joint probability distribution in the new manipulated causal system differs, if at all, from the original distribution before the manipulation.

[28] This scenario will work only if the governor is not functioning!

[29] Readers knowledgeable about tractors will have to assume that the governor of the tractor is defective.

[30] Given the typical assumptions of the statistical test used, and assuming that he is not in the presence of an unusual event.

[31] This statement must be tempered on account of practical problems involving statistical power.

3

Sewall Wright, path analysis and d-separation

3.1 A bit of history

> The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated with well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.
>
> The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.

So begins Sewall Wright's 1921 paper (Wright 1921), in which he describes his ‘method of path coefficients’. In fact, he invented this method while still in graduate school (Provine 1986) and had even used it without presenting its formal description in a paper published the previous year (Wright 1920). The 1920 paper used his new method to describe and measure the direct and indirect causal relationships that he had proposed to explain the patterns of inheritance of different colour patterns in guinea pigs. The paper came complete with a path diagram – i.e. a causal graph – in which actual drawings of the colour patterns of guinea pig coats were used instead of variable names.

Wright was one of the most influential evolutionary biologists of the twentieth century, being one of the founders of population genetics and intimately involved in the modern synthesis of evolutionary theory and genetics. Despite these other impressive accomplishments, Wright viewed path analysis as one of his more important scientific contributions, and continued to publish on the subject right up to his death (Wright 1984). The method was described by his biographer (Provine 1986) as ‘the quantitative backbone of his work in evolutionary theory’. His method of path coefficients is the intellectual predecessor of all the methods described in this book. It is therefore especially ironic that path analysis – the ‘backbone’ of his work in evolutionary theory – has been almost completely ignored by biologists.[1]

This chapter has three goals. First, I want to explore why, despite such an illustrious family pedigree, path analysis and causal modelling have been largely ignored by biologists. To do this I will have to delve into the history of biometry at the turn of the twentieth century, but it is important to understand why path analysis was ignored in order to appreciate why its modern incarnation does not deserve such a fate. Next, I want to introduce a new inferential test that allows one to test the causal claims of the path model rather than only ‘measuring the direct influence along each separate path in such a system’. The inferential method described in this chapter is not the first such test. Another inferential test was developed quite independently by sociologists in the early 1970s, based on a statistical technique called maximum-likelihood (ML) estimation. Since that method forms the basis of modern structural equation modelling, I will postpone its explanation until the next chapter. Finally, I will present some published biological examples of path analysis and apply the new inferential test to them.

3.2 Why Wright's method of path analysis was ignored

I suspect that scientists largely ignored Wright's work on path analysis for two reasons. First, it ran counter to the philosophical and methodological underpinnings of the two main contending schools of statistics at the turn of the twentieth century. Second, it was methodologically incomplete in comparison to R. A. Fisher's statistical methods (Fisher 1925), based on the analysis of variance combined with the randomised experiment, which had appeared at about the same time.

Francis Galton invented the method of correlation. Karl Pearson transformed correlation from a formula into a concept of great scientific importance and championed it as a replacement for the ‘primitive’ notion of causality. Despite Pearson's long-term programme to provide ‘mathematical contributions to the theory of evolution’ (Aldrich 1995), he had little training in biology, especially in its experimental form. He was educated as a mathematician and became interested in the philosophy of science early in his career (Norton 1975). Presumably his interest in heredity and genetics came from his interest in Galton's work on regression, which was itself applied to heredity and eugenics.[2] In 1892 Pearson published a book entitled *The Grammar of Science* (Pearson 1892). In the chapter entitled ‘Cause and effect’ he gave the following definition: ‘Whenever a sequence of perceptions D, E, F, G is invariably preceded by the perception C…, C is said to be the *cause* of D, E, F, G.’ As will become apparent later, his use of the word ‘perceptions’, rather than ‘events’ or ‘variables’ or ‘observations’, was an important part of his phenomenalist philosophy of science. He viewed the relatively new concept of correlation as having immense importance to science and the old notion of causality as so much metaphysical nonsense. In the third edition of his book (Pearson 1911) he even included a section entitled ‘The category of association, as replacing causation’. He had this to say (Pearson 1911: 166):

> The newer, and I think truer, view of the universe is that all existences are associated with a corresponding variation among the existences in a second class. Science has to measure the degree of stringency, or looseness of these concomitant variations. Absolute independence is the conceptual limit at one end to the looseness of the link, absolute dependence is the conceptual limit at the other end to the stringency of the link. The old view of cause and effect tried to subsume the universe under these two conceptual limits to experience – and it could only fail; things are not in our experience either independent or causative. All classes of phenomena are linked together, and the problem in each case is how close is the degree of association.

These words may seem curious to many readers because they express ideas that have mostly disappeared from modern biology. Nonetheless, these ideas dominated the philosophy of science at the beginning of the twentieth century and were at least partially accepted by such eminent scientists as Albert Einstein. Pearson was a convinced phenomenalist and logical positivist.[3] This view of science was expressed by people such as Gustav Kirchhoff, who held that all science can do is discover new connections between phenomena, not discover the 'underlying reasons'. Ernst Mach (Mach 1883), who dedicated one of his books to Pearson, viewed the only proper goal of science as providing economical descriptions of experience by describing a large number of diverse experiences in the form of mathematical formulae. To go beyond this and invoke unobserved entities such as 'atoms' or 'causes' or 'genes' was not science, and such terms had to be removed from its vocabulary. Accordingly, Mach (and Pearson) held that a mature science would express its conclusions as functional – i.e. mathematical – relationships that can summarise and predict direct experience, not as causal links that can explain phenomena (Passmore 1966).

Pearson had thought long and hard about the notion of causality, and he concluded, in accord with British empiricist tradition and the people cited above, that association was all that there was. Causality was an outdated and useless concept. The proper goal of science was simply to measure direct experiences (phenomena) and to economically describe them in the form of mathematical functions. If a scientist could predict the likely values of variable Y after observing the values of variable X then he would have done his job. The more simply and accurately he could do it, the better his science. Referring back to Chapter 2, Pearson did not view the equivalence operator of algebra (=) as an imperfect *translation* of a causal relationship because he did not recognise 'causality' as anything but correlation in the limit.[4] By the time that Wright published his method of path analysis, Pearson's British school of biometry was dominant. One of its fundamental tenets was that 'it is this conception of correlation between two occurrences embracing all relationships from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation' (Pearson 1911: 157).

Given these strong philosophical views, imagine what happened when Wright proposed using the biometrists' tools of correlation and regression…to peek beneath direct observation and deduce systems of causation from systems of correlation! In such an intellectual atmosphere Wright's paper on path analysis was seen as a direct challenge to the biometrists. One has only to read the title ('Correlation and causation') and the introduction of Wright's (1921) paper, cited at the beginning of this chapter, to see how infuriating it must have seemed to the Pearson school.

The pagan had entered the temple, and, like the Macabees, someone had to purify it. The reply came the very next year (Niles 1922). Said Henry Niles: ‘We therefore conclude that philosophically the basis of the method of path coefficients is faulty, while practically the results of applying it where it can be checked prove to be wholly unreliable.’ Although he found fault in some of Wright's formulae (which were, in fact, correct) the bulk of Niles’ scathing criticism was openly philosophical: ‘“Causation” has been popularly used to express the condition of association, when applied to natural phenomena. There is no philosophical basis for giving it a wider meaning than partial or absolute association. In no case has it been proved that there is an inherent necessity in the laws of nature. Causation is correlation…’ (Niles 1922).

Any Mendelian geneticist during that time – of whom Wright was one – would have accepted as self-evident that a mere correlation between parent and offspring told nothing about the mechanisms of inheritance. Therefore, concluded these biologists, a series of correlations between traits of an organism told nothing of how these traits interacted biologically or evolutionarily.[5] The biometricians could never have disentangled the genetic rules determining colour inheritance in guinea pigs, which Wright was working on at the time, simply by using correlations or regressions. Even if distinguishing causation from correlation appeared philosophically ‘faulty’ to the biometricians, Wright and the other Mendelian geneticists were experimentalists for whom statements such as ‘causation is correlation’ would have seemed equally absurd. For Wright, his method of path analysis was not a statistical *test* based on standard formulae such as correlation or regression. Rather, his path coefficients were interpretative parameters for measuring direct and indirect causal effects based on a causal system that had already been determined. His method was a statistical translation, a mathematical analogue, of a biological system obeying asymmetric causal relationships.

As the fates would have it, path analysis soon found itself embroiled in a second heresy. Three years after Wright's ‘Correlation and causation’ paper, Fisher published his *Statistical Methods for Research Workers* (1925). Fisher certainly viewed correlation as distinct from causation. For him the distinction was so profound that he developed an entire theory of experimental design to separate the two. He viewed randomisation and experimental control as the only reliable way of obtaining causal knowledge. Later in his life Fisher even wrote an entire book criticising the research that identified tobacco smoking as a cause of cancer on the basis that such evidence was not based on randomised trials (Fisher 1959).[6] I have already described the assumptions linking causality and probability distributions, unstated by Fisher but needed to infer causation from a randomised experiment, as well as the limitations of these assumptions, when studying different attributes of organisms. Despite these limitations, Fisher's methods had one important advantage over Wright's path analysis: they allowed one to rigorously test causal hypotheses, while path analysis could only estimate the direct and indirect causal effects *assuming* that the causal relationships were correct.

Mulaik (1986) has described these two dominant schools of statistics in the twentieth century. His phenomenalist and empiricist school starts with Pearson. Examples of the statistical methods of this school were correlation, regression,[7] common-factor and principal component analyses. The purpose of these methods was primarily, as Mach directed, to provide an economical description of experience by describing a large number of diverse experiences in the form of mathematical formulae. The second school was the realist school, begun by Fisher. This second school emphasised the analysis of variance, experimental design based on the randomised experiment and the hypothetico-deductive method. These Fisherian methods were not designed to provide functional relationships but, rather, to ensure the conditions under which causal relationships could be reliably distinguished from non-causal relationships.

In hindsight, then, it seems that path analysis simply appeared at the wrong time. It did not fit into either of the two dominant schools of statistics and it contained elements that were objectionable to both. The phenomenalist school of Pearson disliked Wright's notion that one *should* distinguish ‘causes’ from correlations. The realist school of Fisher disliked Wright's notion that one *could* study causes by looking at correlations. Professional statisticians therefore ignored it. Biologists found Fisher's methods, complete with inferential tests of significance, more useful and conceptually easier to grasp, and so biologists ignored path analysis too. A statistical method viewed as central to the work of one of the most influential evolutionary biologists of the twentieth century was largely ignored by biologists.

3.3 D-sep tests

Wright's method of path analysis was so completely ignored by biologists that most biometry texts do not even mention it. Those that do (Li 1975; Sokal and Rohlf 1981) describe it as Wright originally presented it, without even mentioning that it was reformulated by others, primarily economists and social scientists, such that it permitted inferential tests of the causal hypothesis and allowed one to include unmeasured (or ‘latent’) variables. The main weakness of Wright's method – that it required one to assume the causal structure rather that being able to test it – had been corrected by 1970 (Jöreskog 1970), but biologists are mostly unaware of this.

Two different ways of testing causal models will be presented in this book. The most common method is called structural equation modelling and is based on maximum-likelihood techniques. This method will be described in Chapters 4 to 7, and it does have a number of advantages when testing models that include variables that cannot be directly observed and measured (so-called *latent* variables) and for which one must rely on observed indicator variables that contain measurement errors. SEM also has some statistical drawbacks. The inferential tests are asymptotic and can therefore require rather large sample sizes. The functional relationships must be linear. Data that are not multivariate normal are difficult to treat.

These drawbacks led me to develop an alternative set of methods that can be used for small sample sizes, non-normally distributed data or non-linear functional relationships (Shipley 2000). Since these methods are derived directly from the notion of d-separation that was described in Chapter 2, I will call these *d-sep* tests. The main disadvantage of d-sep tests is that they are not applicable to causal models that include latent (unmeasured) variables.

The link between causal conditional independence and probabilistic independence, given by d-separation, suggests an intuitive way of testing a causal model: simply list all the d-separation statements that are implied by the causal model and then test each of them using an appropriate test of conditional independence. There are a number of problems with this naïve approach. First, even models with a small number of variables can include a large number of d-separation statements. Second, we need some way of combining all these tests of independence into a single composite test. For instance, if we had a model that implied 100 independent d-separation statements and tested each independently at the traditional 5 per cent significance level then we would expect, on average, five of these tests to reach significance simply due to random sampling fluctuations. Even worse, the d-separation statements in a causal model are almost never completely independent, and so we would not even know what the true overall significance level would be. Each of these problems can be solved.

3.4 Independence of d-separation statements

Given an acyclic[8] causal graph, we can use the d-separation criterion to predict a set of conditional probabilistic independencies that must be true if the causal model is true. However, many of these d-separation statements can be themselves predicted from other d-separation statements and are therefore not independent. Happily, Pearl (1988) describes a simple method of obtaining the minimum number of d-separation statements needed to completely specify the causal graph and proves that this minimum list of d-separation statements is sufficient to predict the entire set of d-separation statements. This minimum set of d-separation statements is called a *basis set*.[9] The basis set is not unique. This method is illustrated with Figure 3.1.

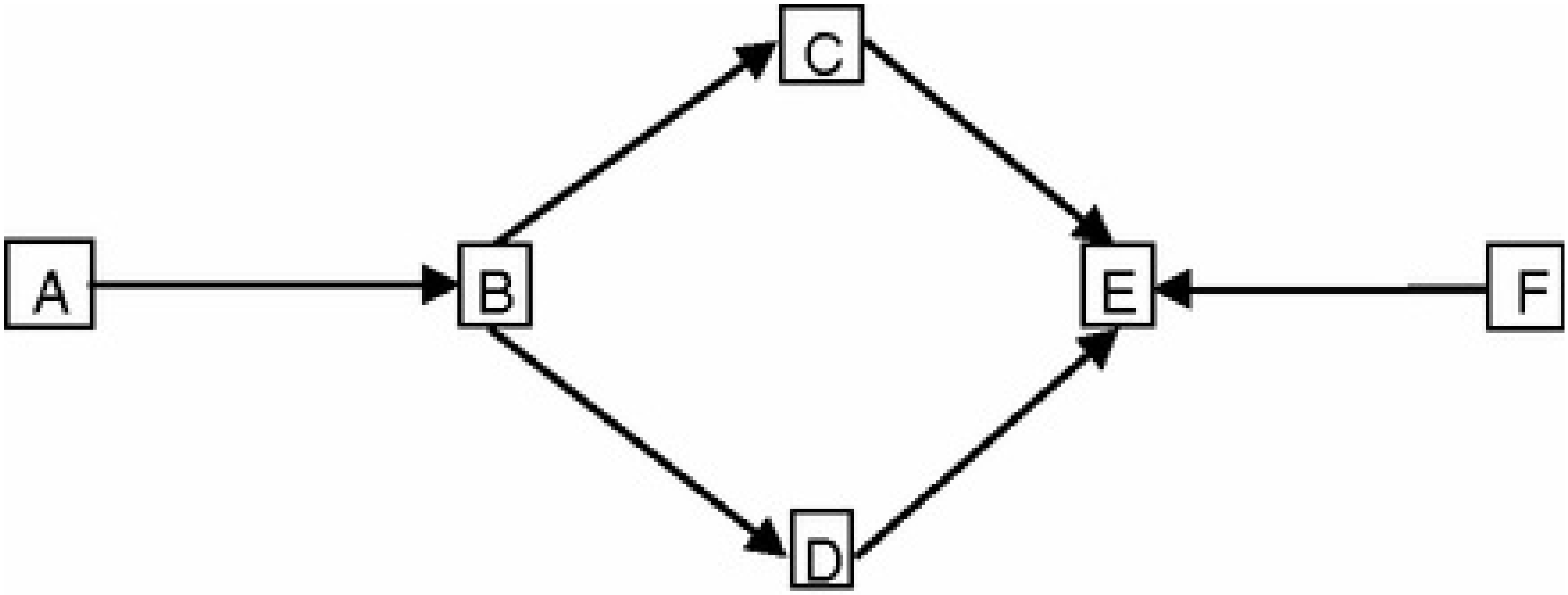


Figure 3.1 A directed acyclic graph involving six variables

To obtain the basis set, the first step is to list each unique pair of non-adjacent vertices – that is, list each pair of variables in the causal model that do not have an arrow between them. So, in Figure 3.1 the list is {(A,C), (A,D), (A,E), (A,F), (B,E), (B,F), (C,D), (C,F), (D,F)}. Pearl's (1988) basis set is given by d-separation statements consisting of each such pair of vertices conditioned on the parents of the vertex having higher causal order. The number of pairs of variables that don't have an arrow between them is always equal to the total number of pairs minus the number of arrows in the causal graph. In general, if there are V variables and A arrows in the causal graph, the number of elements in the basis set will be

$$\frac{V!}{2(V-2)!}-A$$

Unfortunately, the conditional independencies derived from such a basis set are not necessarily mutually independent in finite samples (Shipley 2000). A basis set that does have this property[10] is given by the set of unique pairs of non-adjacent vertices, of which each pair is conditioned on the set of causal parents of both (Shipley 2000). Remember that an exogenous variable has no parents, so the set of 'parents' of such a variable is empty (such an empty set is written {ϕ}). The second step in getting the basis set that will be used in the inferential test, described next, is to list all causal parents of each vertex in the pair. Using Figure 3.1 and the notation for d-separation introduced in Chapter 2,[11] Table 3.1 summarises the d-separation statements that make up the basis set.

Table 3.1 A basis set for the DAG shown in Figure 3.1 along with the implied d-separation statements

Non-adjacent variables	Parent variables of either non-adjacent variable	D-separation statement
A, C	B	$A \perp\!\!\!\perp C \mid B$
A, D	B	$A \perp\!\!\!\perp D \mid B$
A, E	C, D, F	$A \perp\!\!\!\perp E \mid CDF$
A, F	None	$A \perp\!\!\!\perp F$
B, E	A, C, D, F	$B \perp\!\!\!\perp E \mid ACDF$
B, F	A	$B \perp\!\!\!\perp F \mid A$
C, D	B	$C \perp\!\!\!\perp D \mid B$
C, F	B	$C \perp\!\!\!\perp F \mid B$
D, F	B	$D \perp\!\!\!\perp F \mid B$

Each of the d-separation statements in Table 3.1 predicts a (conditional) probabilistic independence. How you test each predicted conditional independence depends on the nature of the variables, and so different d-separation statements in your basis set could be tested with different statistical tests of (conditional) independence. For instance, if the two variables involved in the independence statement are normally and linearly distributed, you could test the hypothesis that the Pearson partial correlation coefficient is zero. Other tests of conditional independence are described below. At this point, assume that you have used tests of independence that are appropriate for the variables involved in each d-separation statement and that you have obtained the exact probability level assuming such independence. By ‘exact’ probability levels, I mean that you cannot simply look at a statistical table and find that the probability is ≤0.05; rather, you must obtain the actual probability level – say, p = 0.036.

Because the conditional independence tests implied by the basis set are mutually independent, we can obtain a composite probability for the entire set using Fisher's test. Since this test seems not to have a name, I have called it Fisher's C (for ‘combined’) test. If there are a total of k independence tests in the basis set, and p_i is the exact probability of the i^{th} test assuming independence, then the test statistic is $C = -2\sum_{i=1}^{k} Ln(p_i)$. If all k independence relationships are true then this statistic will follow a chi-squared distribution with 2k degrees of freedom. This is not an asymptotic test unless you use asymptotic tests for some of the individual independence hypotheses. Furthermore, you can use different statistical tests for different individual independence hypotheses. In this sense, it is a very general test.

3.5 Testing for probabilistic independence

In this section, I want to be more explicit concerning what ‘independence’ and ‘conditional independence’ mean and the different ways that one can test such hypotheses given empirical data. Let's first start with the simplest case: that of unconditional independence.

The difference between the value of a random value X_i and its expected value μ_X is $(X_i–\mu_X)$. Since these differences can be both negative or positive, and we want to know simply the deviation around the expected value, not the direction of the deviation, we can take the square of the difference: $(X_i–\mu_X)^2$. The expected value of this squared difference[12] is the variance:

$$E[(X_i - \mu_X)^2] = E[(X_i - \mu_X)(X_i - \mu_X)]$$

The covariance is simply a generalisation of the variance. If we have two different random variables (X, Y) measured on the same observational units then the covariance between these two variables is defined as $E[(X_i - \mu_X)(Y_i - \mu_Y)]$. If X and Y behave independently of each other then large positive deviations of X from its mean (μ_X) will be just as likely to be paired with large or small, negative or positive, deviations of Y from its mean (μ_Y). These will cancel each other out in the long run (remember, we are envisaging a complete statistical population when we talk about ‘expectation’) and the expected value of the product of these two deviations, $E[(X_i - \mu_X)(Y_i - \mu_Y)]$, will be zero. Therefore, the probabilistic independence of X and Y implies a population zero covariance.[13] If X and Y tend to behave similarly, increasing or decreasing together, then large positive values of X will often be paired with large positive values of Y and large negative values of X will often be paired with large negative values of Y. In such cases, the covariance will be large and positive. If X and Y tend to behave in opposite ways, the covariance between them will be negative.

A Pearson correlation coefficient is simply a standardised covariance. Neither a variance nor a covariance has any upper or lower bounds. Changing the units of measurement (say, from metres to millimetres) will change both the variance and the covariance. If we divide the covariance between two variables by the product of their variances (taking the square root of this product in order to ensure that the range goes from +1 to –1) then we obtain a Pearson correlation coefficient. Box 3.1 summarises these points.

Box 3.1 Variances, covariances and correlations

Population variance (sigma^2, σ^2) of a random variable X:

$$E[(X - \mu_X)^2]$$

Variance (s^2) of a random variable X from a sample of size n:

$$\frac{\sum_i (X_i - \bar{X})^2}{n - 1}$$

Population covariance (sigma$_{XY}$, σ_{XY}) between two random variables X, Y:

$$E[(X - \mu_X)(Y - \mu_Y)]$$

Covariance (s_{XY}) between two random variables X, Y from a sample of size n:

$$\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Population Pearson correlation (rho$_{XY}$, ρ_{XY}) between two random variables, X, Y:

$$\frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

Pearson correlation coefficient (r_{XY}) between two random variables, X, Y from a sample of size n:

$$\frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$$

Variances and covariances can be obtained via the cov() function in R:

```
cov(x, y = NULL, use = “everything”,
method = c(“pearson”, “kendall”, “spearman”))
```

The closely related function cor() gives the correlations; specifiying the method to be ‘pearson’ gives the Pearson correlations:

```
cor(x, y = NULL, use = “everything”,
method = c(“pearson”, “kendall”, “spearman”))
```

The formulae in Box 3.1 are valid so long as both X and Y are random variables. If we want to conduct an inferential test of independence using these formulae, we have to pay attention to the probability distributions of X and Y and the form of the relationship between them in case they are not independent. Different assumptions concerning these points require different statistical methods.

Case 1: X and Y are both normally distributed and any relationship between them is linear.

Tests of the independence of X and Y involving this set of assumptions are treated in any introductory statistics book. First, one can transform the Pearson correlation coefficient so that it follows Student's t-distribution. If X and Y, sampled randomly and measured on n units, are independent (so, the null hypothesis is that ρ = 0) then the following transformation will follow a Student's t-distribution[14] with n–2 degrees of freedom:

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

This test is exact. So long as you have at least three independent observations then you can test for the independence of X and Y.[15] The R function cor.test, with method = ‘pearson’, performs this test.

It is also possible to transform a Pearson correlation coefficient so that it asymptotically follows a standard normal distribution – i.e. a normal distribution with a mean of zero and a variance of 1. For sample sizes of at least 50 (and approximately even for sample sizes as low as 25) one can use Fisher's *Z*-transform:

$$z = 0.5\sqrt{n-3}\,\ln\left(\frac{1+r}{1-r}\right)$$

If X and Y are independent then the probability of z can be obtained from a standard normal distribution. Finally, one can use Hotelling's (1953) transformation,[16] which is acceptable for sample sizes as low as 10:

$$z = \sqrt{(n-1)}\left[0.5\ln\left(\frac{1+r}{1-r}\right) - \frac{1.5\ln\left(\frac{1+r}{1-r}\right)+r}{4(n-1)}\right]$$

Case 2: X and Y are continuous but not normally distributed and any relationship between them is only monotonic.

If X or Y are not normally distributed and any relationship between them is not linear but is monotonic[17] then we can use Spearman's correlation coefficient. Although there exist statistical tables giving probability levels for Spearman's correlation coefficient, one can use exactly the same formulae as for Pearson's correlation coefficient so long as the sample size is greater than 10 (Sokal and Rohlf 1981).

The first step is to convert X and Y to their ranks. In other words, sort each value of X from smallest to largest and replace the actual values of X by its order in the rank; the smallest number becomes 1, the second smallest number becomes 2, and so on. Do the same thing for Y. Now that you have converted each X and each Y to its rank, you can simply put these numbers into the formula for a Pearson's correlation coefficient and test as before.

One complication is when there are ties. Spearman's coefficient assumes that the underlying values of X and Y are continuous, not discrete. Given such an assumption, equal values of X (or Y) will occur only because of limitations in measurement. To correct for such ties, first sort the values ignoring ties, and then replace the ranks of tied values by the mean rank of these tied values. Box 3.2 gives an example of the calculation of a Spearman rank correlation coefficient.

Box 3.2 Spearman's rank correlation coefficient

Here are 10 simulated pairs of values and the accompanying scatterplot (Figure 3.2). The X values were drawn from a uniform distribution and rounded to the nearest unit. The Y values were drawn from the following equation, $Y_i = X_i^{0.2} + \beta(5, 1)$, where the random component is drawn from a beta distribution with shape parameters of 5 and 1.

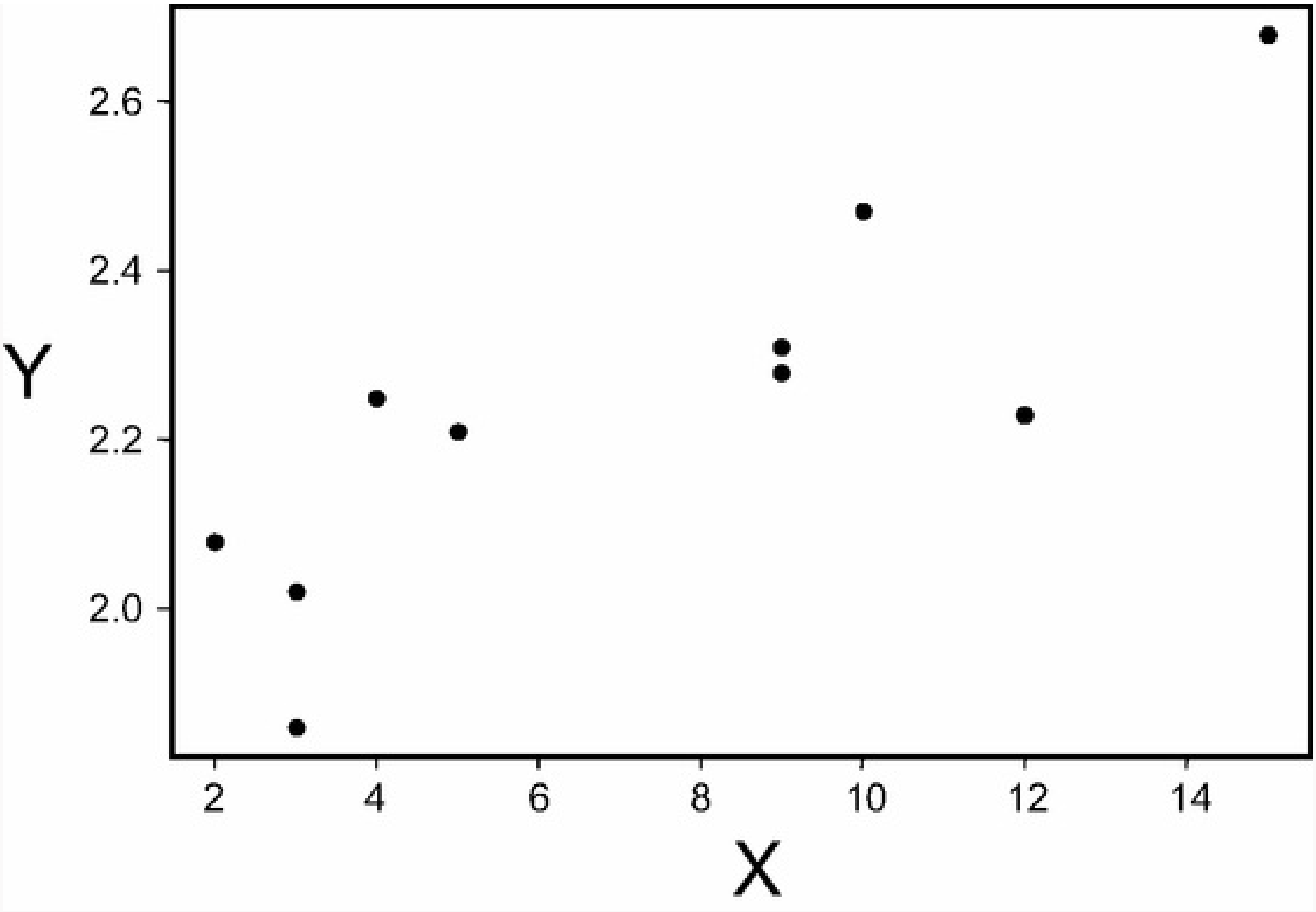


Figure 3.2 A scatterplot of randomly generated pairs of values from a bivariate non-normal distribution and possessing a non-linear monotonic relationship

Values of X, Y and their ranks

X	Y	Rank X	Rank Y	Rank X	Rank Y
2	2.08	1	3	1	3
3	2.02	2	2	2.5	2
15	2.68	10	10	10	10
10	2.47	8	9	8	9
5	2.21	5	4	5	4
12	2.23	9	5	9	5
3	1.86	3	1	2.5	1

4	2.25	4	6	4	6
9	2.31	6	8	6.5	8
9	2.28	7	7	6.5	7

In the above table, X and Y are the original values. Columns 3 and 4 of the table are the ranks of X and Y before correcting for ties (the underlined values). Columns 5 and 6 are the ranks after correcting for the two pairs of ties values of X (there were two values of 3 and two values of 9). To calculate the Spearman rank correlation coefficient of X and Y, simply use the values in columns 5 and 6 and enter them into the formula for the Pearson's correlation coefficient. In the above example the Spearman rank correlation coefficient is 0.726. Assuming that X and Y are independent in the statistical population, we can convert this to a standard normal variate using Hotelling's z-transform, giving a value of 2.47. This value has a probability under the null hypothesis of 0.014.

The R function cor.test, with method = ‘spearman’, performs this test.

Case 3: X and Y are continuous and any relationship between them is not even monotonic.

This case applies when the relationship between X and Y might have a very complicated form, with X and Y being positively related in some parts of the range and negatively related in other parts, and therefore when neither a Pearson nor a Spearman correlation can be applied. This situation requires more computationally demanding methods, including form-free regression and permutation tests. Each of these topics is dealt with much more fully in other publications but will be intuitively introduced here because these notions are needed for the analogous case in conditional independence. Form-free regression is a vast topic, which includes kernel smoothers, cubic-spline smoothers (Wahba 1991) and local (loess) smoothers (Cleveland and Devlin 1988; Cleveland, Devlin and Grosse 1988; Cleveland, Grosse and Shyu 1992). Collectively, these methods form the basis of generalised additive models[18] (Hastie and Tibshirani 1990). Permutation tests for association are described by Good (1993; 1994).

3.6 Permutation tests of independence

To begin, consider a simple linear regression of Y on X, where both are random variables. The correlation between X and Y is the same as the correlation between the observed value of Y and the predicted value of Y given X – that is, E[Y|X]. To test for an association between X and Y in this regression context we need to do three things. First, we have to estimate the predicted values of Y for each value of X. For linear regression we simply obtain the slope and intercept to get these values, and in the general case we would use form-free regression methods. Second, we need to calculate a measure of the association between the observed and predicted values of Y; we can use a Pearson correlation coefficient, a Spearman correlation coefficient or any of a large number of other measures that can be found in the statistical literature. Finally, we need to know the probability of having observed such a value when, in fact, X and Y really are independent. This is where a permutation test comes in handy.

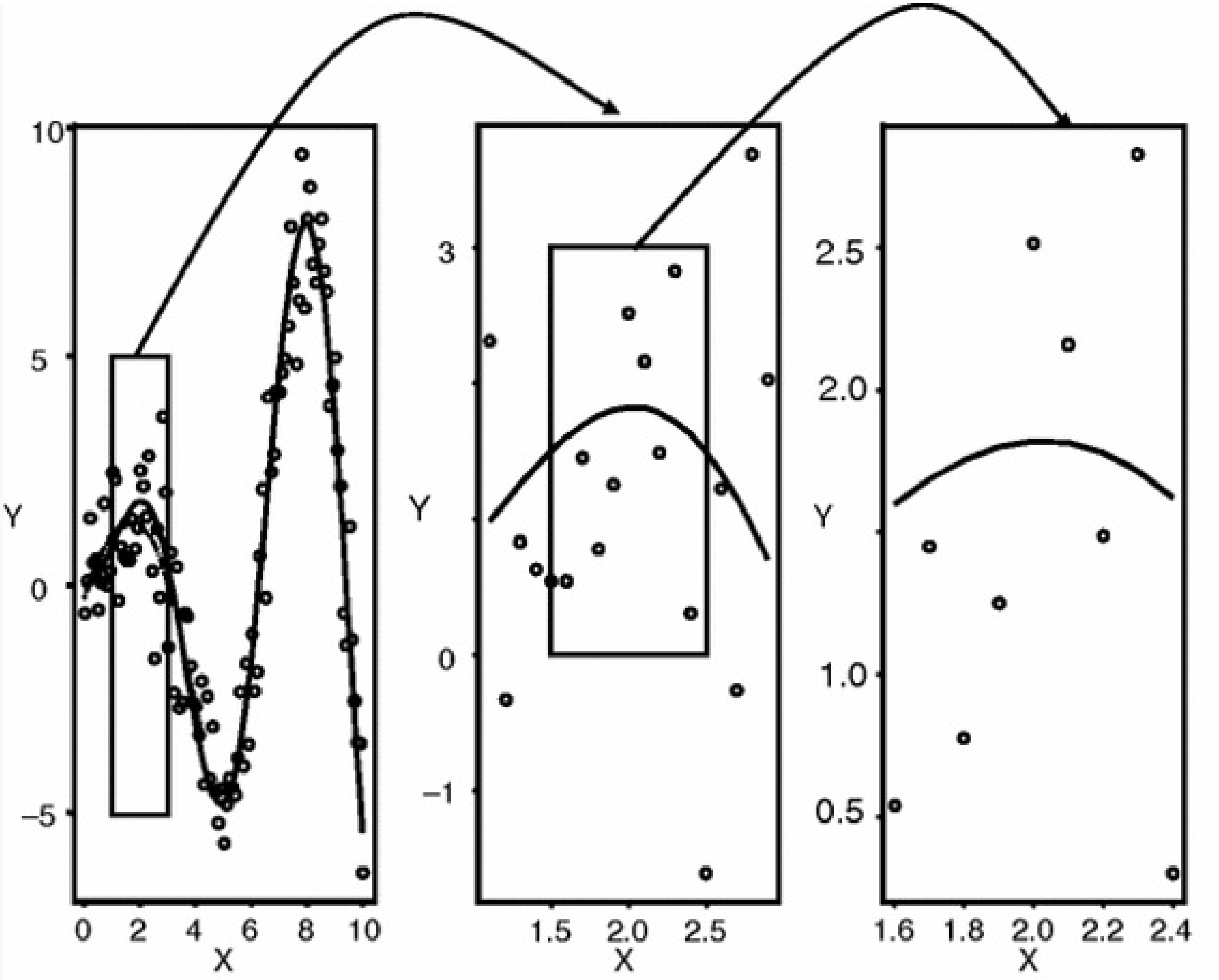
Remember the definition of probabilistic independence given in Chapter 2. We know that, if X and Y are independent, the probability of observing any particular value of Y is the same regardless of whether we know the value of X. In other words, any value of X is just as likely to be paired with any other value of Y as with the particular Y that we happen to observe. The permutation test works by making this true in our data. After calculating our measure of association in our data, we randomly rearrange the values of X and/or Y using a random number generator. In this new randomly mixed 'data set' the values of X and Y really are independent, because we forced them to be so; we have literally forced our null hypothesis of independence to be true, and the value of the association between X and Y is due only to chance. We do this a very large number of times until we have generated an empirical frequency distribution of our measure of association.[19] The exact number of times that we randomly permute our data will depend on the true probability level of our actual data and the accuracy that we want to obtain in our probability estimate. Manly (1997) shows how to determine this number, but it is typically between 1,000 and 10,000 times. On modern computers this will appear instantaneously unless the intermediate calculations are intensive. The last step is to count the proportion of times that we observe at least as large a value of association within the permuted data sets, or its absolute value for a two-tailed test, as we actually observed in our original data. Box 3.3 gives an example of this permutation procedure.

3.7 Form-free regression

The first graph in Box 3.3 (Figure 3.3(a)) shows a highly non-linear relationship between X and Y, and it is unlikely that we would be able to deduce the actual function that generated these data.[20] On the other hand, if we concentrate on smaller and smaller sections of the graph (Figures 3.3(b) and (c)), the relationship becomes simpler and simpler. The basic insight of form-free regression methods is that even complicated functions can be quite well approximated by simple linear, quadratic or cubic functions in the neighbourhood of a given value of X. Within such a neighbourhood, shown by the boxes in the graphs of Box 3.3, we can use these simpler functions to calculate the expected value of Y at that particular value of X. We then go on to the next value of X, move the neighbourhood so that it is centred around this new value of X and calculate the expected value of the new Y, and so on. In this way, we do not actually estimate a parametric function predicting Y over the entire range of X, but we do get very good estimates of the predicted values of Y given each unique value of X. To obtain the predicted values of Y given X, we use weighted regression (linear, quadratic or cubic) where each (X, Y) pair in the data set is weighted according to its distance from the value of X around which the neighbourhood is centred. In local, or loess,[21] regression the neighbourhood size can be chosen according to different criteria, such as minimising the residual sum of squares, and the weights are chosen on the basis of the tricube weight

Box 3.3 Loess regression

The following three graphs (Figure 3.3) show a simulated data set generated from a complicated non-linear function (the solid line of Figure 3.3(a)) along with a loess regression (the broken line) using a local quadratic fit and a neighbourhood size of one-half the range of X. Figure 3.3(b) shows the same complicated non-linear function in the range of 1 to 3 of the X values and Figure 3.3(c) shows this in the range of 1.5 to 2.5 of the X values.



(a) This graph shows a highly non-linear function (the solid line) between X and Y and the loess fit (the dotted line, mostly superimposed on the solid line)

(b) This graph reproduces the small rectangle in (a)

(c) This graph reproduces the small rectangle in (b)

Figure 3.3 A simulated data set

The loess regression (the broken line in Figure 3.3(a)) doesn't actually give a parametric function linking Y to X, but it does give the predicted value of Y for each unique value of X – that is, it gives the sample estimate of E[Y|X]; the solid and broken lines in the figure completely overlap except in the range of X = 2. To estimate a permutation probability of the non-linear correlation of X and Y, we can first calculate the Pearson correlation coefficient between the observed Y values (the circles in the figure) and the predicted values of Y given

X (the loess estimates). In this example, r = 0.956. If we don't want to assume any particular probability distribution for the residuals then we can generate a permutation frequency distribution for the correlation coefficient. To do this, we randomly permute the order of the observed Y values (or the predicted values; it doesn't matter which) to get a 'new' set of Y^* values and recalculate the Pearson correlation coefficient between Y^* and $E[Y^*|X]$. The following histogram (Figure 3.4) shows the relative frequency of the Pearson correlation coefficient in 5,000 such permutations; the arrow indicates the value of the observed Pearson correlation coefficient. None of the 5,000 permutation data sets had a Pearson correlation coefficient whose absolute value was at least 0.956. Since the residuals were actually generated from a unit normal distribution, we can calculate the probability of observing a value of 0.956 with 101 observations. It is approximately 1×10^{39}.

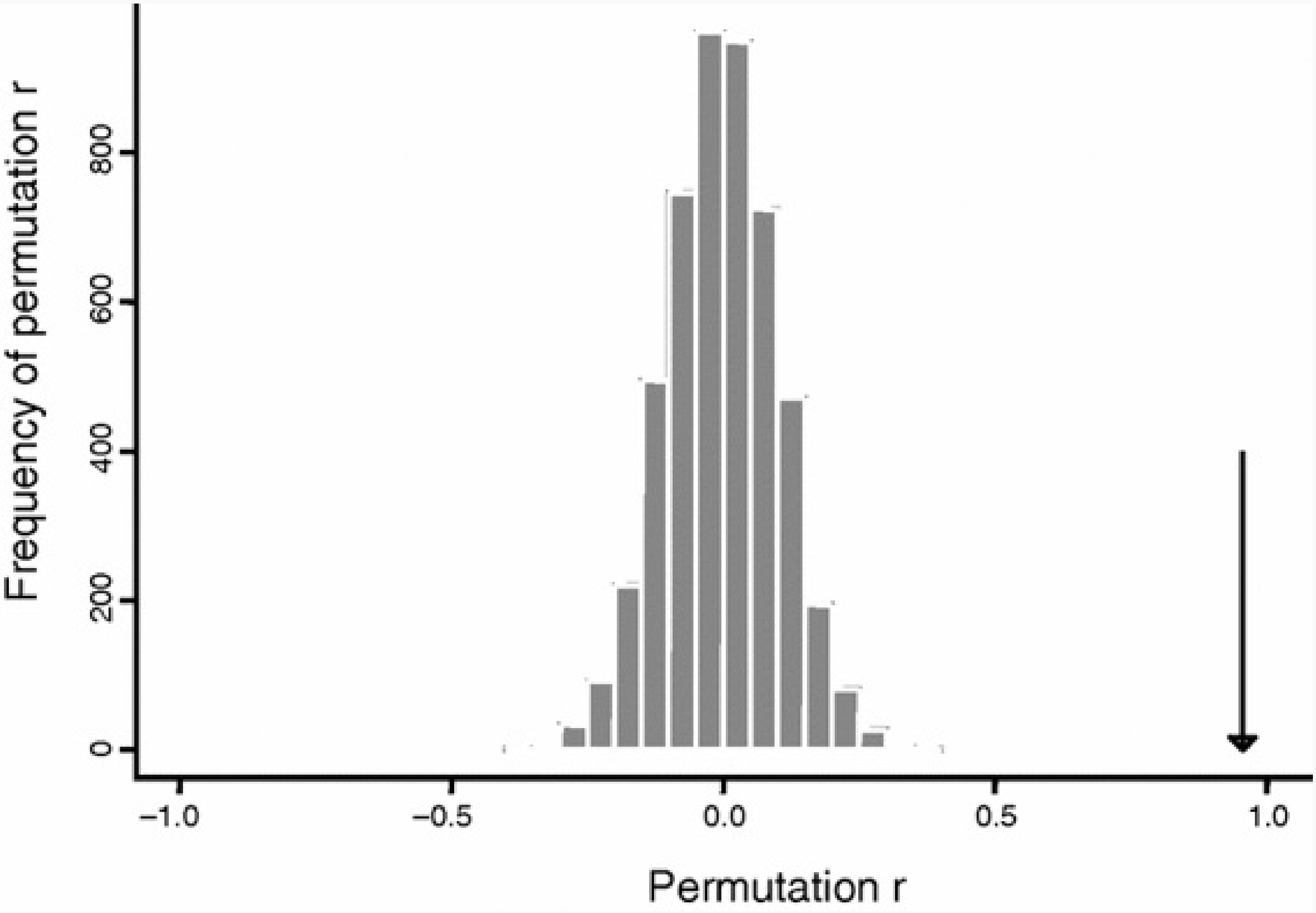


Figure 3.4 The frequency distribution of the Pearson correlation coefficient in 5,000 random permutations of the simulated data set involving the observed Y values and the predicted loess values

Note: The arrow shows the observed Pearson correlation in the original simulated data set.

function. Shipley and Hunt (1996) describe this in more detail in the context of plant growth rates.

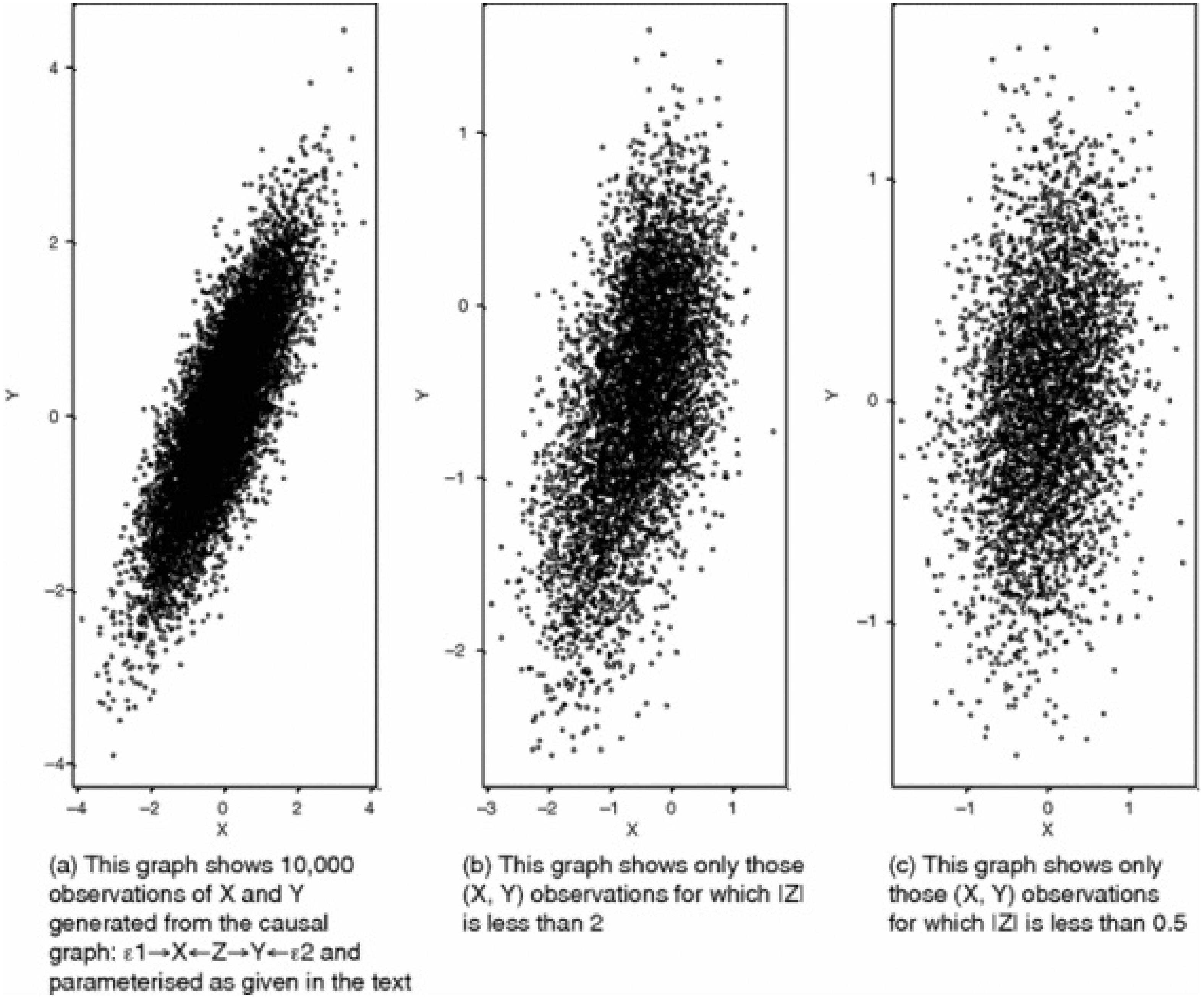
3.8 Conditional independence

So far we have been talking about unconditional independence – that is, the independence of two variables without regard to the behaviour of any other variables. Such unconditional independence is implied by two variables in a causal graph that are d-separated without conditioning on any other variable. D-separation upon conditioning implies *conditional* independence. The notion of conditional independence seems paradoxical to many people. How can two variables be dependent, even highly correlated, and still be independent upon conditioning on some other set of variables?

Consider the following causal graph: ε1→X←Z→Y←ε2. Does it seem equally paradoxical if I say that X and Y will behave similarly due to the common causal effect of Z, but that they will no longer behave similarly if I prevent Z from changing? If Z doesn't change then the only changes in X and Y will come from the changes in ε1 and ε2, and these two variables are d-separated and therefore unconditionally independent. A moment's reflection will convince you that if Z is allowed to change (vary) then both X and Y will change as well in a systematic fashion, since they are both responding to Z. If the variables in the causal graph are random then the correlation between X and Y will be attributable to the fact that both share common variance due to Z. If we restrict the variance in Z more and more then X and Y will share a smaller and smaller amount of common variance. In the limit, if we prevent Z from changing at all then X and Y will no longer share any common variance; the only variation in X and Y will come from the independent error variables ε1 and ε2, and so X and Y will then be independent. In such a case we would be comparing values of X and Y when Z is constant. This is the intuitive meaning of conditional independence. To illustrate, I generated 10,000 independent sets of ε1, X, Z, Y and ε2 according to the following generating equations:

$$\begin{aligned}
\varepsilon 1 &= N(0, 1 - 0.9^2) \\
\varepsilon 2 &= N(0, 1 - 0.9^2) \\
Z &= N(0, 1) \\
Y &= 0.9Z + \varepsilon 1 \\
X &= 0.9Z + \varepsilon 2
\end{aligned}$$

Since X, Y and Z are all unit normal variables, the population correlations are $\rho_{X,Z} = 0.9$, $\rho_{Y,Z} = 0.9$ and $\rho_{X,Y} = 0.81$. Figure 3.5 shows three scatterplots. Notice that X and Y are highly correlated even though neither X nor Y is a cause of the other. Figure 3.5(a) shows the relationship between X and Y when no restrictions are placed on the variance of Z. The sample correlation between X and Y in this graph is 0.8016, compared to the population value of 0.81. Figure 3.5(b) plots only those values of X and Y for which the value of Z is between –2 and 2, which means that the variance of Z is restricted by just a little bit. The sample correlation between X and Y has been decreased slightly to 0.7591. Figure 3.5(c) plots those values of X and Y for which the value of Z is between –0.5 and 0.5, thus restricting the variance of Z much more. The sample correlation between X and Y is now only 0.2294. Clearly, the degree of association between X and Y is decreasing as Z is prevented more and more from varying.



(a) This graph shows 10,000 observations of X and Y generated from the causal graph: ε1→X←Z→Y←ε2 and parameterised as given in the text

(b) This graph shows only those (X, Y) observations for which |Z| is less than 2

(c) This graph shows only those (X, Y) observations for which |Z| is less than 0.5

Figure 3.5 Three scatterplots generated from the causal graph ε1→X←Z→Y←ε2

If we calculate the correlation between X and Y as we restrict the variation in Z more and more, we can get an idea of what happens to the correlation between X and Y in the limit when the variance of Z is zero. This limit is the correlation between X and Y when Z is fixed (or 'conditioned') to a constant value; this is called the *partial* correlation between X and Y, conditional on Z, and it is written $\rho_{XY.Z}$ or $\rho_{XY|Z}$. Figure 3.6 plots the sample correlation between X and Y as Z is progressively restricted in its variance.

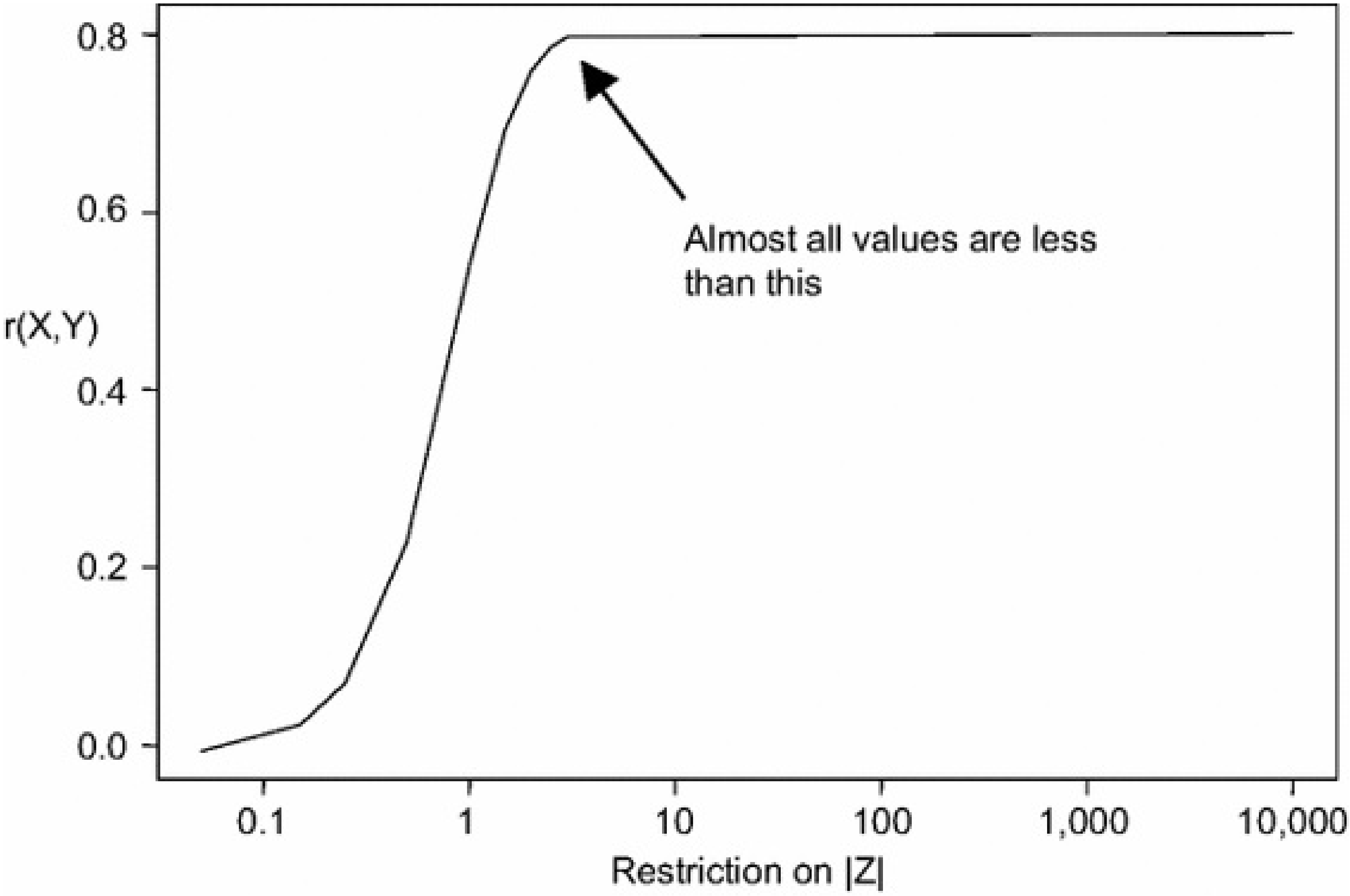


Figure 3.6 The Pearson correlation coefficient between X and Y in the data shown in Figure 3.5 when the absolute value of Z is restricted to various degrees

Note: The limiting value of the correlation coefficient when |Z| is restricted to a constant value is the partial correlation between X and Y.

As expected, as the range of Z around its mean (zero) becomes smaller and smaller, the correlation between X and Y also becomes smaller and approaches zero. Given the causal graph that governed these data we know that X and Y are not unconditionally d-separated and therefore are not unconditionally independent. However, X and Y are d-separated given Z, and therefore X and Y are independent conditional on Z.

If we remember that a regression of X on Z gives the expected value of X conditional on Z then the residuals around this regression are the values of X for fixed values of Z. This gives us another way of visualising the partial correlation of X and Y conditional on Z: it is the correlation between the residuals of X, conditional on Z, and the residuals of Y, conditional on Z. If I regress, in turn, each of X and Y on Z in the above example and calculate the correlation coefficient between the residuals of these two regressions, I get a value of –0.0060.

This view of a conditional independence provides us with a very general method of testing for it. If X and Y are predicted to be d-separated given some other set of variables **Q** = {A, B, C,…}, regress (perhaps using form-free regression) each of X and Y on the set **Q** and then test for the independence of the residuals, using, if you want, any of the methods of testing unconditional independence described above. If the residuals are normally distributed and linearly related then you can use the test for Pearson correlations. If the residuals appear, at most, to have a monotonic relationship then you can use the test for a Spearman correlation. If the residuals have a more complicated pattern then you can use one of the non-parametric smoothing techniques available, followed a permutation test. The only difference is that you have to reduce the degrees of freedom in the tests by the number of variables in the conditioning set.

If your statistical program can invert a matrix then there are faster ways of calculating partial Pearson or Spearman correlations. These are explained in Box 3.4. The pcor() function in the ggm library calculates Pearson or Spearman partial correlations. If your matrix or data frame is called ‘my.dat’ then a typical call to this function would be

```
pcor(u,cov(my.data))
```

. Here, u is a vector giving the variable names or column numbers of the variables in your data for which you want to calculate the partial correlation. The first two names or numbers in u index the variables whose partial correlation you want and any remaining names or numbers in u index the conditioning variables.

Box 3.4 Calculating partial covariances and correlations

Given a sample covariance matrix **S**, the inverse of this matrix is called the *concentration* matrix, **C**. The negative of the off-diagonal elements c_{ij} give the partial covariance between variables i and j, conditional on (holding constant) all the other variables included in the matrix. This gives an easy way of estimating partial covariances and partial correlations of any order. To get the partial covariance between variables X and Y conditional on a set of other variables **Q**, simply create a covariance matrix in which the only variables are X, Y and the remaining variables in **Q**. Invert the matrix, and then this partial covariance is the negative of the element in the row pertaining to X and the column pertaining to Y – namely $-c_{XY}$. The partial correlation between X and Y is given by

$$r_{X,Y|Q} = \frac{-c_{XY}}{\sqrt{c_{XX} \cdot c_{YY}}}$$

The partial correlation between two variables conditioned on n other variables is said to be a *partial correlation of order n*. The unconditional correlation coefficient is simply a partial correlation of order 0. Some texts give recursion formulae for partial correlations of various orders, although partials of higher orders are very tedious to calculate by such means. For instance, the formula for a partial correlation of order 1 between X and Y, conditional on Z, is

$$\rho_{X,Y|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{\left(1 - \rho_{XZ}^2\right)\left(1 - \rho_{YZ}^2\right)}}$$

As an example, consider the following causal graph: W→X→Z→Y. 100 independent (W,X,Y,Z) observations were generated according to structural equations with all path coefficients equal to 0.5 and the variances of all four variables equal to 1.0. Here is the sample covariance matrix:

$$\begin{array}{ccccc} & W & X & Y & Z \\ W & 1.43347870 & -0.75265627 & -0.06269845 & 0.10179918 \\ X & -0.75265627 & 1.52762094 & -0.53911722 & -0.03777874 \\ Y & -0.06269845 & -0.53911722 & 1.71116716 & -0.90033856 \\ Z & 0.10179918 & -0.03777874 & -0.90033856 & 1.73196991 \end{array}$$

The inverse of the matrix (rounded to the nearest 100th) obtained by extracting only the elements of the covariance matrix pertaining to W, X and Y is

$$\begin{array}{cccc} & W & X & Y \\ W & 1.43 & -0.75 & -0.01 \\ X & -0.75 & 1.53 & -0.56 \\ Y & -0.01 & -0.56 & 1.24 \end{array}$$

The partial correlation between W and Y, conditional on X, is

$$r_{WY|X} = \frac{(-1) - 0.01}{\sqrt{1.43 \cdot 1.24}} = 0.0075$$

The same method can be used to obtain partial Spearman partial correlations, by simply ranking the variables as described in Box 3.2 and then proceeding in the same way as for Pearson partial correlations.

3.9 Spearman partial correlations

This next section presents some Monte Carlo results to explore the degree to which the sampling distribution of Spearman partial correlations, after appropriate transformation, follows either a standard normal or a Student's t distribution. This section is not necessary to understand the application of d-sep tests for path models, only to justify the use of Spearman partial correlations in testing for conditional independence. If you don't need to know this then you can skip this section.

There has been remarkably little published in the primary literature concerning inferential tests related to non-parametric conditional independence.[22] It is known that the expected values of a first-order partial Kendall or Spearman partial correlations need not be strictly zero even when two variables are conditionally independent given the third (Shirahata 1980; Korn 1984). On the other hand, Conover and Iman (1981) recommend the use of partial Spearman correlations for most practical cases in which the relationships between the variables are at least monotonic. A Spearman partial correlation is simply a Pearson partial correlation applied to the ranks of the variables in question. Therefore, the conditional independence of non-normally distributed variables with non-linear, but monotonic, functional relationships between the variables can be tested with Spearman's partial rank correlation coefficient simply by ranking each variable (and correcting for ties, as described in Box 3.2) and then applying the same inferential tests as for Pearson partial correlations. For instance, if one accepts Conover and Iman's (1981) recommendations then a Spearman partial rank correlation will be approximately distributed as a standard normal variate when z-transformed.

How robust is this recommendation? To explore this question, Table 3.2 presents the results of some Monte Carlo simulations to determine the effects of sample size, the distributional form of the variables and the effect of non-linearity on the sampling distribution of the z-transformed Spearman partial correlation coefficient. The random components of the generating equations (ε_i) were drawn from four different probability distributions: normal, gamma, beta or binomial. I chose the shape parameters of the gamma and beta distributions to produce different degrees of skew and kurtosis. Gamma(λ = 1) is a negative exponential distribution. Gamma(λ = 5) is an asymmetric distribution with a long right tail. Beta(1,1) is a uniform distribution, Beta(1,5) is a highly asymmetric distribution with a long right tail and Beta(5,1) is a highly asymmetric distribution with a long left tail. The final (discrete) probability distribution was symmetric with an expected value of 2 and had ordered states of X = 0, 1, 2, 3 or 4; these were generated from a binomial distribution of the form $C(5,X)0.5^X 0.5^{1-}$

X. Random numbers were generated using the random number generators given by Press et al. (1986). The generating equations were of the form

$$X_1 = \varepsilon_1$$
$$X_i = \alpha_i X_{(i-1)}^{\beta_i} + \varepsilon_i; \quad i > 1$$

These generating equations are based on a causal chain ($X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \ldots$) with sufficient variables (3, 4 or 5) to produce zero partial associations of orders 1 to 3. When β_i equals 1.0 the relationships between the variables are linear and when β_i is different from 1.0 then the relationships between the variables are non-linear but monotonic. The results in Table 3.2 are based on models with β_i = 1 (linear) and 0.5 (non-linear) but other values give similar results. All the simulation results in Table 3.2 are based on 1,000 independent simulated data sets. In interpreting Table 3.2, remember that the z-transformed Spearman partial correlations should be approximately distributed as a standard normal variate whose population mean is zero, whose population standard deviation is 1.0 and whose two-tailed 95 per cent limit is 1.96.

Table 3.2 Results of a Monte Carlo study of the distribution of z-transformed Spearman partial correlations

Distribution of ε_i	Sample size	Order of partial	Linear/non-linear	Mean of z	Standard deviation of z	Two-tailed 95% quantile	Theoretical probability
Normal	25	1	L	0.08	1.03	2.04	0.04
Normal	50	1	L	0.08	0.97	2.01	0.05
Normal	400	1	L	0.08	1.04	2.16	0.03
Normal	50	2	L	0.03	0.99	1.86	0.06
Normal	50	3	L	−0.07	1.00	1.85	0.06
Gamma(1)	25	3	L	0.01	1.05	2.09	0.04
Gamma(1)	50	3	L	−0.02	0.96	1.82	0.07
Gamma(1)	50	3	NL	0.03	0.96	2.02	0.04
Gamma(5)	50	3	NL	−0.07	0.99	1.93	0.05

Beta(1,1)	50	3	L	−0.02	0.99	2.00	0.05
Beta(1,1)	50	3	NL	0.03	1.02	2.08	0.04
Beta(1,5)	50	3	NL	0.03	1.02	2.08	0.04
Beta(5,1)	50	3	NL	−0.05	0.99	1.78	0.07
Beta(5,1)	400	3	NL	0.00	1.02	2.01	0.04
Binomial	50	3	NL	0.01	0.99	1.95	0.05

Notes: Four different distributional types were simulated for the random components. The sample size was the number of observations per simulated data set. Linear (L) and non-linear (NL) functional relationships were used. The empirical mean, the standard deviation and the two-tailed 95 per cent limits of 1,000 simulated data sets are shown.

Generally, the sampling distribution of the z-transformed Spearman rank partial correlations is a very good approximation of a standard normal distribution. In fact, the only significant deviation from a standard normal distribution (based on a Kolmogorov–Smirnov test) was observed for the ranks of normally distributed variables, for which one would not normally use a Spearman partial correlation. The empirical standard deviations were always close to 1.0 and the empirical means only once differed significantly, though very slightly, from zero at high levels of replication. Approximate 95 per cent confidence intervals for the empirical 0.05 significance level (i.e. the two-tailed 95 per cent quantiles), based on 1,000 simulations, are 0.037 to 0.064 (Manly 1997).

The results of this simulation study support the recommendations of Conover and Iman (1981). These results are also consistent with the theoretical values given by Korn (1984) for the special case of a Spearman first-order partial based on trivariate normal and trivariate lognormal distributions, where the limiting values of the Spearman partial correlation are less than, or equal to, an absolute value of 0.012, thus giving an expected absolute z-score of ≤0.024. Korn (1984) gives a pathological example in which the above procedure will not work even after ranking the data because there is a non-monotonic relationship between the variables; he recommends that one first check[23] to see if the relationship between the ranks are approximately linear before using Spearman partial correlations.

3.10 Seed production in St Lucie cherry

St Lucie cherry (*Prunus mahaleb*) is a small species of tree that is found in the Mediterranean region and that relies on birds for the dispersal of its seeds. As in most plants, seedlings from seeds that are dispersed some distance from the adult are more likely to survive, since they will not be shaded by their own parent or eaten by granivores that are attracted to the parent tree. For species whose seeds can survive the passage through the digestive tract of the dispersing animal, it is also evolutionarily and ecologically advantageous for the fruit to be eaten by the animal, since the seed will be deposited with its own supply of fertiliser. Not all frugivores of St Lucie cherry are useful fruit dispersers. Some birds just consume the pulp, either leaving the naked seed attached to the tree or simply dropping the seed to the ground directly beneath the parent. In order to estimate selection gradients, Jordano (1995) measured six traits of 60 individuals of this species: the canopy projection area (a measure of photosynthetic biomass), average fruit diameter, the number of fruits produced, average seed weight, the number of fruits consumed by birds and the percentage of these consumed fruits that were properly dispersed away from the parent by passage through the gut. Based on five of these variables for which I had data (I was missing the total number of fruits consumed by birds) I proposed the path model shown in Figure 3.7 (Shipley 1997), using the exploratory path models described in Chapter 8.

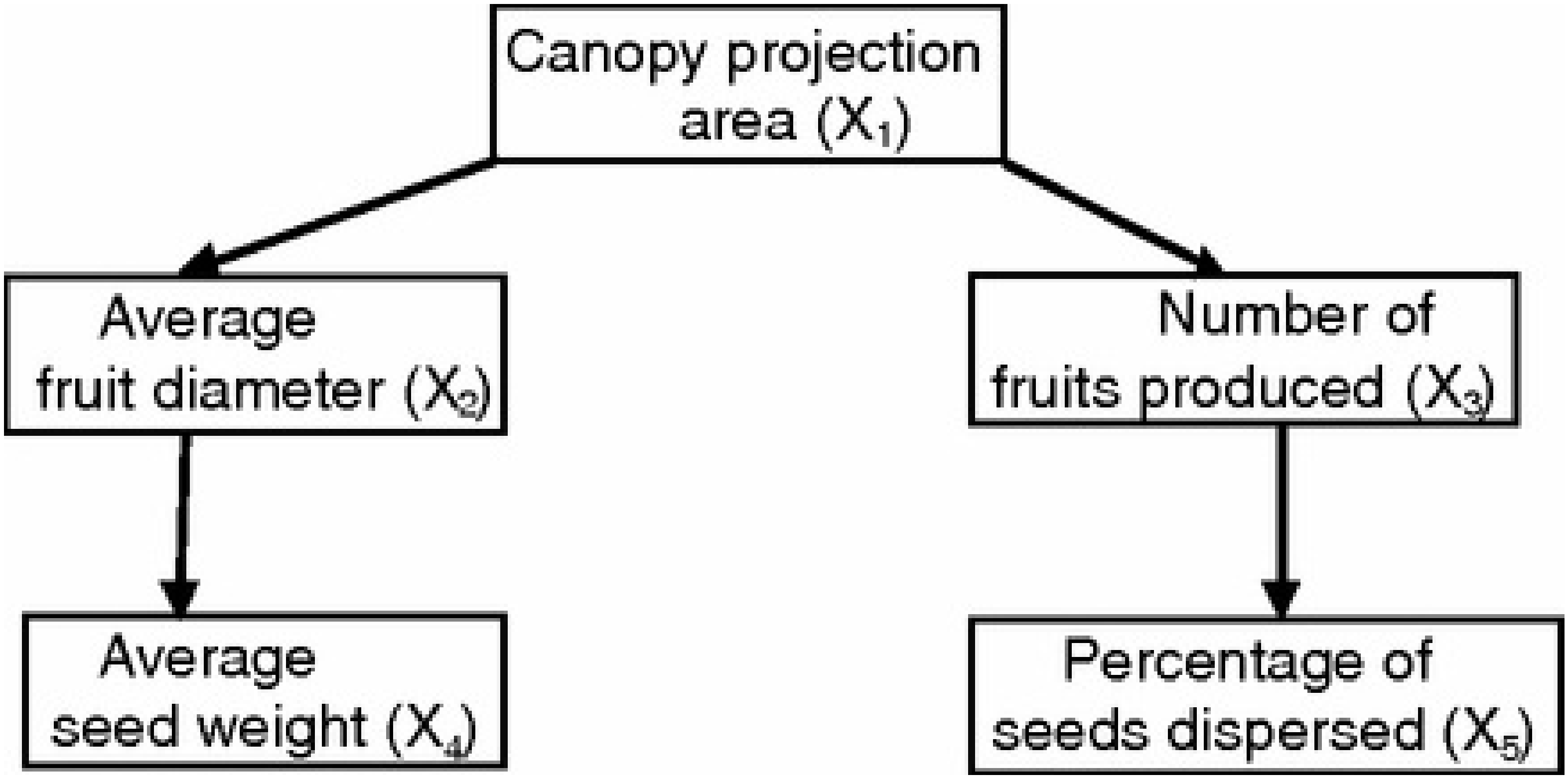

Canopy projection area (X_1)
Average fruit diameter (X_2)
Number of fruits produced (X_3)
Average seed weight (X_4)
Percentage of seeds dispersed (X_5)

Figure 3.7 Proposed causal relationships between five variables related to seed dispersal in St Lucie cherry

We can use this model to illustrate the d-sep test. The first step is to obtain the d-separation statements in the basis set that are implied by the causal graph in Figure 3.7. There are six such statements, since there are five variables and four arrows. Table 3.3 lists these d-separation statements. You can also obtain this basis set using the basiSet() function of the ggm library of R. To do this you must first enter the causal graph shown in Figure 3.7 using the DAG() function that was described in Chapter 2.

```
Figure3.7←DAG(X2~X1,X3~X1,X4~X2,X5~X3)
basiSet(Figure3.7)
```

Table 3.3 The d-separation statements in the basis set of the causal graph shown in Figure 3.7

	Pearson partial correlations		**Spearman partial correlations**	
D-separation statement	**Estimate**	**Probability assuming independence**	**Estimate**	**Probability assuming independence**
$X_4 \perp\!\!\!\perp X_1 \mid X_2$	−0.066	0.617	−0.063	0.635
$X_4 \perp\!\!\!\perp X_3 \mid X_2X_1$	0.142	0.289	0.144	0.279

$X_4 \perp\!\!\!\perp X_5 \mid X_2X_3$	0.004	0.976	0.075	0.574
$X_2 \perp\!\!\!\perp X_3 \mid X_1$	0.021	0.873	0.059	0.655
$X_2 \perp\!\!\!\perp X_5 \mid X_3X_1$	−0.155	0.244	−0.160	0.229
$X_1 \perp\!\!\!\perp X_5 \mid X_3$	0.076	0.565	0.102	0.443

Note: Also shown are the Pearson and Spearman partial correlations that are implied by the d-separation statements. The probabilities, assuming that the population partial correlations are zero, are listed as well.

We next have to decide how to test the independencies that are implied by these six d-separation statements. The original data showed heterogeneity of variance, as often happens with size-related variables, but transforming each variable to its natural logarithm stabilises the variance. Figure 3.8 shows the scatterplot matrix of these ln-transformed data.

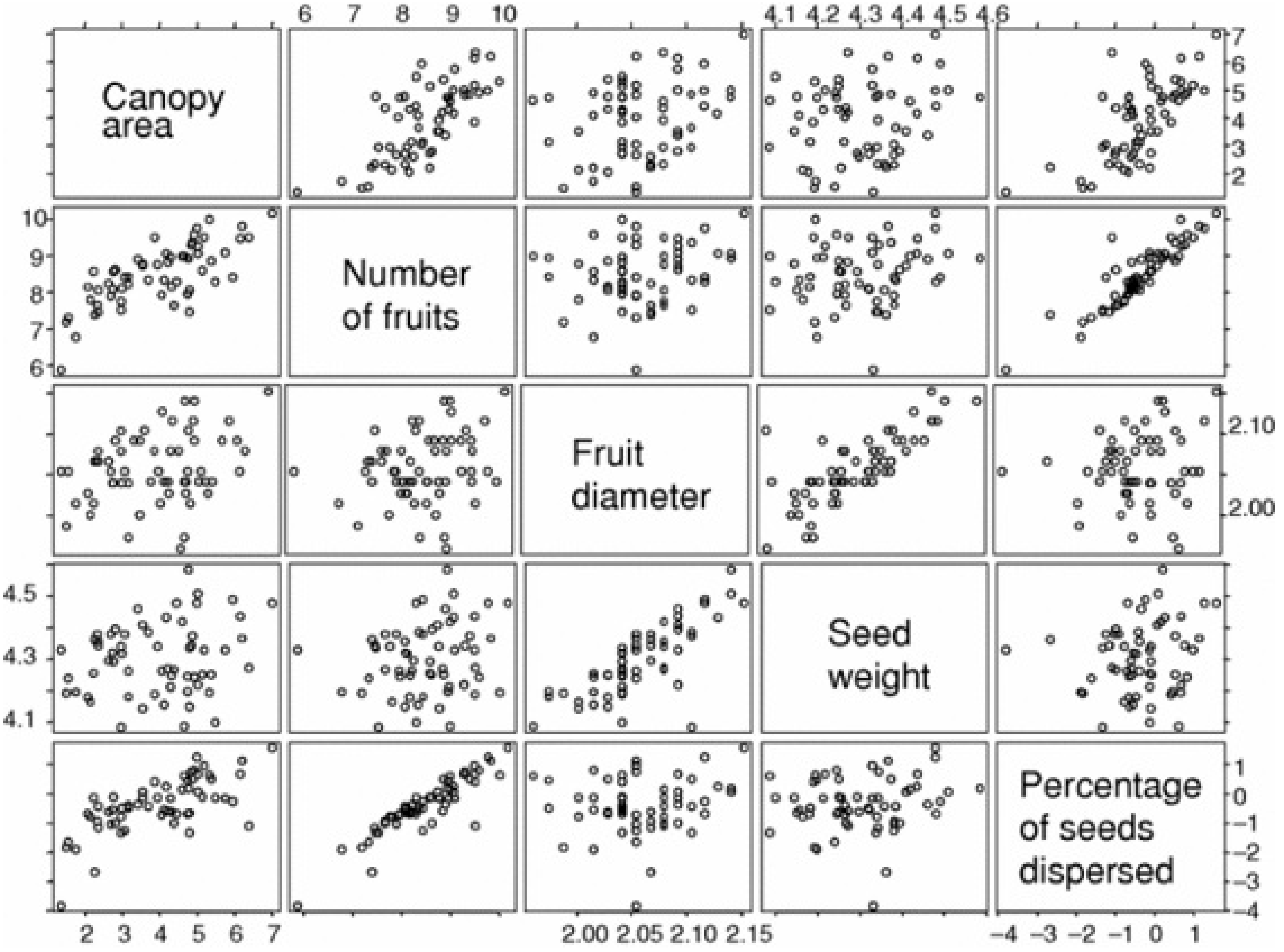


Figure 3.8 Scatterplot matrix of the empirical observations (all variables transformed to their natural logarithms)

Since the relationships appear to be linear and histograms of each variable did not show any obvious deviations from normality, we can test the predicted independencies using Pearson partial correlations. The results[24] are shown in Table 3.3. Fisher's C statistic is 7.73, with 12 degrees of freedom, for an overall probability of 0.806. The difference between the observed and predicted (partial) correlations would occur in about 80 per cent of data sets (in the long run) even if the data really were produced by the causal structure in Figure 3.7. This doesn't mean that the data really were produced by such a causal structure but it does mean that we have no reason to reject it on the basis of the statistical test. If we want to reject it anyway then we will need to produce reasonable doubt. Perhaps the assumption of normality, upon which the test of the Pearson partial correlations is based, was producing incorrect probability estimates. Table 3.3 also lists the Spearman partial correlations. The overall probability of the model (χ^2 = 9.99, 12 df), based on the individual probability levels of these Spearman partial correlations, was 0.616. On the other hand, there are equivalent models that also produce non-significant probability estimates (Shipley 1997), and if any of these equally well-fitting alternative models do not contradict what is known of the biology of these trees then they might constitute reasonable doubt.[25] Equivalent models are explained in Chapter 8.

The original data used by Jordano (1995) was fitted to a latent variable model using maximum likelihood methods.[26] Neither the model chi-square statistic nor the model degrees of freedom were given. It is therefore not possible to judge the fit of that original model,[27] but it is possible to extract those d-separation statements involving only the measured variables available to me from the original latent-variable model. Jordano's published model implies four d-separation statements in the basis set that can be tested: {(canopy projection area $\perp\!\!\!\perp$ average fruit diameter), (canopy projection area $\perp\!\!\!\perp$ average seed weight), (number of fruits produced $\perp\!\!\!\perp$ average fruit diameter), (number of fruits produced $\perp\!\!\!\perp$ average seed weight)}. Combining the null probabilities using Fisher's C test, based on Pearson correlations, gives a probability of 0.005 ($\chi^2 = 21.85$, 8 df). Using Spearman correlations the probability is 0.019 ($\chi^2 = 18.24$). These low probabilities, based on a subset of the original measured variables, provide reasonable doubt concerning Jordano's (1995) model.

3.11 Generalising the d-sep test

These are the steps involved in a d-sep test.

(1) Write down your causal hypothesis in the form of a DAG.

(2) Use the basiSet() function to get the set of d-sep statements forming your basis set, and thus the series of null hypotheses of (conditional) independence implied by them.

(3) Conduct the series of statistical tests of (conditional) independence specified in step (2) using whatever statistical tests are appropriate for each null hypothesis.

(4) Combine the null probabilities, obtained from step (3), using Fisher's C test.

(5) Reject your causal hypothesis if the null probability from Fisher's C test is below your chosen significance level; otherwise, conclude that your data are consistent with your causal hypothesis.

Stated thus, it is clear that the d-sep test is less a statistical ‘test’ than a recipe for creating your own test based on the specific properties of your data. So long as you can (a) specify your causal hypothesis in the form of a DAG, (b) obtain direct measures of each variable in this DAG and (c) perform an appropriate test of (conditional) independence for each d-sep statement then you can create your very own d-sep test.

Once the d-sep test tells you that the data are consistent with the causal process represented in the DAG, you can calculate the path coefficients and related statistics. The details of how to do this will vary depending on the nature of your data but the first step will be the same: you fit a series of structured equations by exactly following the causal structure specified by your DAG. In other words, you identify each effect variable and, for each one, you fit the data to a model in which this effect variable is predicted jointly by its causal parents and by no other variable.

In this section I want to explain how you can use the full power of linear models, generalised linear models, mixed models, non-linear models and generalised additive models to carry out step (3) above. I do not explain how to conduct these various analyses here since each type could (and does) fill an entire book. If you are reading this book then you probably know at least how to fit and interpret a linear model (simple and multiple regression, ANOVA, ANCOVA,[28] etc.).

Consider any d-sep statement of the form $X \perp\!\!\!\perp Y|\mathbf{Q}$, where **Q** is the set of conditioning variables. We can translate this d-sep statement into a statistical model either as $Y \sim f(X) + f(\mathbf{Q}) + \varepsilon$ or as $X \sim f(Y) + f(\mathbf{Q}) + \varepsilon$. The notation $f(\ \cdot\ )$ simply means 'some function of' whatever variables are inside the brackets and ε is a random variable following some probability distribution. Depending on the details of the data this function may be linear or non-linear. Depending on the details of the data the random variable could be normally distributed or follow some other probability distribution, it could be a scalar or a vector of values, and so on. Depending on these differences you would then choose a statistical model appropriate for these details. Should variable X or Y be the dependent variable (i.e. the variable whose probability distribution we need to know)? It doesn't matter for what follows, and so you can choose the one that is easier to model. Once you decide the nature of your dependent variable (i.e. what probability distribution you should assume for ε), the nature of the predictor variables (i.e. whether they are continuous, discrete counts or ordered/unordered factors) and the form of the functions linking the dependent and predictor variables, you will have decided which type of statistical model to use.

The basis set of your DAG, obtained from the basiSet() function, consists of a series of d-sep claims, each having the general form $X \perp\!\!\!\perp Y|\mathbf{Q}$. For each of these d-sep claims you will choose an appropriate statistical model based on the considerations discussed above. In order to test each d-sep claim, which is a null hypothesis stating that variables X and Y are independent conditional on the set **Q** of other variables, you will then obtain the probability that the partial slope of X in this model is zero in the statistical population. Why? Because a partial slope of X that is zero is equivalent to saying that X and Y are independent conditional on the variables in **Q**, which is the null hypothesis specified by the d-sep claim. As an added complication, the order in which the independent variables (i.e. X and **Q**) are added to the model can change the null hypothesis that is associated with a zero partial slope in some types of models, and the null hypothesis that we want to test is that X is independent of Y after taking into account the variables in **Q**. Therefore, a good rule of thumb is to always enter the **Q** variables into the model before X. I have explained how to do this (Shipley 2009) in the context of nested data having both normally distributed and binomially distributed variables, along with an example. However, you can adapt this to many different contexts as long as your causal model is a DAG and as long as you can test the (conditional) independence claims implied by your basis set with an appropriate statistical test that provides null probabilities.

An example

The leaves of most flowering plants are photosynthetic organs. Since carbon fixation is so central to the survival of plants, one might expect there to be a tight integration of leaf form and physiology to provide for this necessary function. However, land plants face a dilemma. They need to keep their tissues turgid, but these humid tissues find themselves surrounded by air or soil that is not saturated with water. The leaves (and other tissues) are protected by a cuticle to prevent dehydration. Unfortunately, this severely restricts not only the diffusion of water vapour but also other gases, especially the carbon dioxide that is required for photosynthesis, from diffusing into the leaves. The production of stomates is the evolutionary solution to this problem. Stomates are small openings on the surface of the leaves through which gases can diffuse, and the size of the stomatal openings is controlled by guard cells.

As soon as the stomates begin to open, carbon dioxide begins to diffuse from the outside air into the intercellular spaces of the leaf through a process of passive diffusion. Since the leaf is photosynthesising, carbon dioxide is being removed from the intercellular spaces, creating a diffusion gradient. However, the air inside the leaf is always saturated with water vapour. As soon as the stomates begin to open, this water vapour also begins to diffuse out of the leaf, since the outside air is not saturated with water. In essence, the leaf has to accept a loss (water) in order to effect a gain (carbon). Cowan and Farquhar (1977) have proposed a theoretical model of stomatal regulation to predict how the leaf should control its stomates in order to maximise carbon gain relative to water loss. The basic insight of this model is that the leaf should restrict carbon fixation below the maximum level, because when the internal CO_2 level in the leaf reaches a certain level the main carboxylating enzyme (ribulose-1,5-bisphosphate carboxylase/oxygenase: RuBisCO) becomes saturated, and further increases in carbon fixation require the regeneration of adenosine triphosphate (ATP) from the light reaction of photosynthesis. The second stage results in a greatly reduced rate of increase of carbon fixation per increase in the internal CO_2 concentration, but the rate of water loss continues at its former rate. Thus, Cowan and Farquhar's principal insight is that the leaf should maintain the intercellular CO_2 concentration at the break point between RuBisCO limitation and ribulose-1,5-bisphosphate (RuP_2) regeneration limitation so that the carboxylating capacity and the capacity to regenerate RuBisCO are co-limiting.

Based on these theoretical notions, Martin Lechowicz and I (Shipley and Lechowicz 2000) have proposed a path model based on five variables:

(1) specific leaf mass (SLM: leaf dry mass divided by leaf area, g/m^2);

(2) leaf organic nitrogen concentration (mmol/m^2);

(3) stomatal conductance to water (mmol/m^2/s);

(4) net photosynthetic rate (µmol/m^2/s); and

(5) internal CO_2 concentration (µl/l).

The proposed model is shown in Figure 3.9. Our data were the mean values from 40 herbaceous species typical of wetland environments.

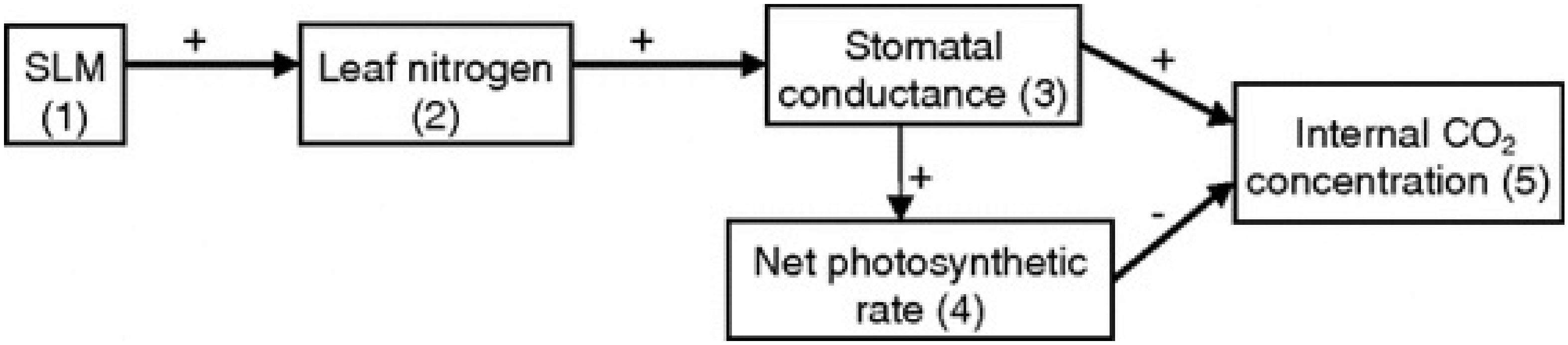

SLM (1)
+
Leaf nitrogen (2)
+
Stomatal conductance (3)
+
+
Internal CO_2 concentration (5)
-
Net photosynthetic rate (4)

Figure 3.9 Proposed causal relationships between five variables related to interspecific leaf morphology and gas exchange

There are five outliers in the data in relation to the internal CO_2 concentration. These are 'C_4' species. The other 35 species are C_3 species. C_4 species have an additional metabolic pathway in which atmospheric carbon is first fixed by phosphoenolpyruvate (PEP) carboxylase in the mesophyll cells to form malate or aspartate. This molecule, a 4-carbon acid, is then transferred into bundle-sheath cells deeper in the leaf. Here these C_4 acids are decarboxylated, and the freed carbon dioxide enters the normal Calvin cycle of the dark reaction of photosynthesis. An advantage of C_4 photosynthesis is that plants exhibiting it are able to absorb CO_2 strongly from a lower concentration of CO_2 within the leaf. They can do this without RuBisCO acting as an oxygenase, rather than a carboxylase, under conditions of low CO_2 and high O_2. This means that C_4 plants do not exhibit the wasteful process of photorespiration under conditions of high illumination and low water availability. Because of this, they are able to maintain high rates of photosynthesis even when the stomates are nearly closed. The basis set implied by the model in Figure 3.9, along with the relevant statistics, is summarised in Table 3.4.

Table 3.4 The d-separation statements on the basis implied by the model in Figure 3.9

	Both C_3 and C_4 species				Only C_3 species			
	Pearson		Spearman		Pearson		Spearman	
D-sep	**r**	**p(r)**	**r**	**p(r)**	**r**	**p(r)**	**r**	**p(r)**
1 ⫫ 3|2	−0.286	0.0777	−0.234	0.1523	−0.298	0.0871	−0.226	0.1986
1 ⫫ 4|3	0.165	0.3163	0.217	0.1841	0.109	0.5392	0.188	0.2860
1 ⫫ 5|3,4	0.035	0.8328	0.099	0.5560	0.043	0.8139	0.215	0.2303
2 ⫫ 4|1,3	−0.092	0.5837	−0.069	0.6809	0.160	0.3743	0.156	0.3870
2 ⫫ 5|1,3,4	0.262	0.1169	0.058	0.7327	−0.079	0.6678	−0.006	0.9758
Fisher's C	χ^2 = 13.15, 10 df, p = 0.216		χ^2 = 9.713, 10 df, p = 0.466		χ^2 = 9.301, 10 df, p = 0.503		χ^2 = 10.621, 10 df, p = 0.388	

Notes: Also shown are the Pearson and Spearman partial correlations and their two-tailed probabilities. Results are shown for the full data set of 40 species and for the 35 species of C_3 species only. Numbers refer to the variables shown in Figure 3.9.

There is no strong evidence for any deviation of the data from the predicted correlational shadow, as given by the d-separation statements. However, a reasonable alternative model would be that the leaf nitrogen content, which is primarily due to enzymes related to photosynthesis, directly causes the net photosynthetic rate. In other words, what if Cowan and Farquhar's (1977) model of stomatal regulation is wrong, and the leaf is regulating its stomates to maximise the net rate of CO_2 fixation independently of water loss? In this case, the observed rate of stomatal conductance would be a consequence of the net photosynthetic rate rather than its cause and the net photosynthetic rate would be directly caused by leaf nitrogen content. We can test this alternative model too, and Table 3.5 summarises the results.

Table 3.5 The d-separation statements implied by an alternative model

	Both C_3 and C_4 species				Only C_3 species			
	Pearson		Spearman		Pearson		Spearman	
D-sep	r	p(r)	r	p(r)	r	p(r)	r	p(r)
$1 \perp\!\!\!\perp 3\|2$	−0.286	0.0777	−0.234	0.1523	−0.298	0.0871	−0.226	0.1986
$1 \perp\!\!\!\perp 3\|4$	0.286	0.0777	0.279	0.0853	0.221	0.2092	0.1723	0.3298
$1 \perp\!\!\!\perp 5\|3,4$	0.035	0.8328	0.099	0.5560	0.043	0.8139	0.215	0.2303
$2 \perp\!\!\!\perp 3\|1,4$	0.599	7×10^{-5}	0.569	2×10^{-4}	0.371	0.0338	0.339	0.0541
$2 \perp\!\!\!\perp 5\|1,3,4$	0.262	0.1169	0.058	0.7327	−0.079	0.6678	−0.006	0.9758
Fisher's C:	$\chi^2 = 33.96$, 10 df, p = 0.0002		$\chi^2 = 27.59$, 10 df, p = 0.0021		$\chi^2 = 16.00$, 10 df, p = 0.1000		$\chi^2 = 14.27$, 10 df, p = 0.161	

Notes: In this alternative the model in Figure 3.9 is changed to make leaf nitrogen cause the net photosynthetic rate, which then causes the observed rate of stomatal conductance, along with the Pearson and Spearman partial correlations and their two-tailed probabilities. Results are shown for the full data set of 40 species and for the 35 species of C_3 species only.

This alternative model is clearly rejected when both the C_3 and C_4 species are analysed together, since there are only about two out of 10,000 changes of observing such a large difference by chance. This lack of fit is coming from the predicted independence between leaf nitrogen level (2) and stomatal conductance (3), conditioned jointly on specific leaf mass (1) and net photosynthetic rate (4). This, of course, is the critical distinction between the path model in Figure 3.9 and the alternative model. When looking only at the C_3 species, the alternative model does not have a large degree of lack of fit, though the critical prediction still shows a reasonably large lack of fit ($r_{2,3|1,4} = 0.371$, $p = 0.0338$) and is always poorer than that provided by the structure shown in Figure 3.9.

Because of such results, and other reasons described in the original reference, I prefer the causal structure shown in Figure 3.9. However, such a conclusion must remain tentative. After all, the conclusion is based on only 40 species, and a larger sample size might detect some more subtle lack of fit that was too small to be found in the present data set.

Given the model in Figure 3.9, and given that we have not been able to reject it, we can now fit the path equations. Although Sewall Wright's original method was based on standardised variables, I prefer to use the original variables, because the variables each have well-established units of measurement. The least-squares regression equations, using only the C_3 species, are shown below. The residual variation is indicated by $N(0,\sigma)$.

$$\mathrm{Ln}(\%\ \mathrm{nitrogen}) = 0.78+0.90\mathrm{Ln}(\mathrm{SLM})+N(0,0.243),\ R = 0.85$$

$$\mathrm{Ln}(\mathrm{conductance})= -6.60+1.15\mathrm{Ln}(\%\ \mathrm{nitrogen})+N(0,0.56),\ R = 0.69$$

$$\mathrm{Ln}(\mathrm{photo}) = 3.08+0.55\mathrm{Ln}(\mathrm{conductance})+N(0,0.31),\ R = 0.81$$

$$\mathrm{Ln}(CO_2\ \mathrm{internal}) = 6.42+0.14\mathrm{Ln}(\mathrm{conductance})–0.1\mathrm{Ln}(\mathrm{photo})+N(0,0.04),\ R = 0.77$$

Each of the slopes is significant at a level below 10^{-4} and the sign of each is in the predicted direction. With these path equations we can begin to simulate how the entire suite of leaf traits would change if we change the specific leaf mass (the exogenous variable in this model) or if we observe species with different specific leaf masses. We get the functional relationships by back-converting the variables in the equations from their logarithms. Of course, each of these variables may also change with changing environmental conditions. By including these environmental variables we could generate the response surfaces across which the suite of leaf traits would move as the environment changes.

A suggestion when proposing your own causal models

The following is a common experience among educators; I call it the ‘real-world paradox’. The students arrive on their first day to class with a pre-existing, ‘naïve’ understanding of some phenomenon and how to deal with it. The teacher then spends several weeks in a classroom setting in which he or she teaches the theoretical concepts of a discipline and how to apply these theoretical concepts in practice. These theoretical concepts, and their application, differ from the original, ‘naïve’ understanding of the students. The students demonstrate, through their course work, that they understand these new concepts and how to apply them. After the course is finished and the final mark is posted, the students then return to the ‘real world’ and are presented with the same type of problem that they have already mastered in the classroom. They then proceed to ignore their classroom training and fall back on their ‘naïve’ methods. This occurs because people often associate ideas and methods with the context in which they were obtained. The new theoretical ideas and methods were obtained in a very particular and structured context (a classroom with a professor looking on), which is a context that is very different from the ‘real’ world. Because the original naïve concepts were obtained in the real world, these are the default ones that the students use. The same thing might happen to you when you finish reading this book.

I have noticed that beginners to causal modelling, even after having the notions presented so far in this book explained to them, often make a common mistake when it comes time to leave the classroom and do causal modelling on their own data. They will show me a DAG with lots of arrows and, when I ask why they placed an arrow going from variable x to variable y, they will respond that this is because previous studies have shown that x and y are correlated. This is, of course, the wrong answer! An arrow between x and y (x→y) in a DAG means that changing x would cause a change in y even if we could experimentally hold constant all other variables in our DAG. It is irrelevant that we cannot actually carry out such a manipulation. The wrong answer of our student occurs because their previous real-world experience of data analysis involved purely observational models involving associations between variables.

To help you avoid this mistake, I suggest that you take some time before reading further and close this book. Go to wherever you do your ‘real’ work and place in front of you a picture of your research organism or site and then begin writing down a DAG representing the causal structure linking the variables that interest you. For each pair of variables (x, y), say out loud: ‘I don't care if x and y are associated. If I could physically hold constant all other variables in my model except for x and y, but not any variables outside my model, and then if I could manipulate x, would the value of variable y change or not?' If the answer is 'Yes' then add an arrow going from x to y. If the answer is 'No' then don't add an arrow. It doesn't matter if this experimental manipulation can be performed in practice. The point is to equate an arrow in a DAG with a hypothetical controlled experiment, not with a statistical association.

[1] This was written in the first edition. Path analysis and structural equation modelling are now more common.

[2] Galton published his *Hereditary Genius* in 1869 (Galton 1869), in which he studied the 'natural ability' of men (women were presumably not worth discussing). He was interested in 'those qualities of intellect and disposition, which urge and qualify a man to perform acts that lead to reputation'. He concluded that '[those] men who achieve eminence, and those who are naturally capable, are, to a large extent, identical'. Lest we judge Galton and Pearson too harshly, remember that such views were considered almost self-evident at the time. Charles Darwin is reputed to have said of Galton's book: 'I do not think I ever in my life read anything more interesting and original… a memorable work' (Forrest 1974).

[3] It is more accurate to say that his ideas were a forerunner to logical positivism.

[4] And yet, citing David Hume, Pearson did accept that associations could be time-ordered from past to future. Nowhere in his writings have I found him express unease that such asymmetries could not be expressed by the equivalence operator.

[5] Pearson was strongly opposed to Mendelism, and, according to Norton (1975), this opposition was based on his philosophy of science; Mendelians insisted on using unobserved entities ('genes') and forces ('causation').

[6] Fisher was a smoker. I wonder what he would have thought if, because of a random number, he had been assigned to the 'non-smoker' group in a clinical trial?

[7] Regression based on least squares had, of course, been developed well before Pearson by people such as Carl Friedrich Gauss and had been based on a more explicit causal assumption that the independent variable plus independent measurement errors were the causes of the dependent variable. This distinction lives on under the guise of type I and type II regression.

[8] This restriction will be partly removed later. Remember that d-separation also implies probabilistic independence in cyclic causal models in which all variables are discrete and in

cyclic causal models in which functional relationships are linear.

[9] Let **S** be the set of d-separation facts (and therefore the set of conditional independence relationships) that are implied by a directed acyclic graph. A basis set **B** for **S** is a set of d-separation facts that implies, using the laws of probability, all other elements of S, but no proper subset of **B** sustains such implications.

[10] I call this the ‘union’ basis set. I describe a function from the GGM library of R (basiSet) in a few more pages.

[11] In other words, $X \perp\!\!\!\perp Y|\mathbf{Q}$ means that vertex X is d-separated from vertex Y, given the set of vertices **Q**.

[12] The formula to estimate this in a sample is given in Box 3.1.

[13] But not the converse! One can have a zero covariance among variables that are still dependent if the relationship is non-linear.

[14] For partial correlations, described below, one simply replaces r with the value of the partial correlation coefficient, and the numerator (n–2) becomes (n–2–p), where p is the number of conditioning variables.

[15] Of course, with so few observations you would have so little statistical power that only very strong associations would be detected.

[16] Both Fisher's and Hotelling's transformations can be used to test null hypotheses in which ρ equals a value different from zero. This useful property allows one to compute confidence intervals around the Pearson correlation coefficient.

[17] A non-monotonic relationship is one in which X increases with increasing Y over part of the range and decreases with increasing Y over another part of the range. If you think that a graph of X and Y has hills and valleys then the relationship is non-monotonic.

[18] Note that smooth.spline() and loess() implement cubic-spline smoothers and loess smoothers in R, respectively. The mgcv package of R includes many functions for generalised additive models.

[19] For small samples one can generate all unique permutations of the data. The use of random permutations, described here, is generally applicable, and the estimated probabilities converge on the true probabilities as the number of random permutations increases.

[20] The actual function was: Y = Xsin(X)+ε, where the error term comes from a unit normal distribution.

[21] The word 'loess' comes from the geological term 'loess', which is a deposit of fine clay or silt along a river valley. I suppose that this evokes the image of a very wavy surface that traces the form of the underlying geological formation. At least some statisticians have a sense of the poetic.

[22] Kendall and Gibbons (1990) briefly discuss Spearman and Kendall partial correlations and provide a table of significance values for first-order Kendall partial correlations for small sample sizes.

[23] This can be done by simply plotting the scatterplots of the ranked data.

[24] For the case of normally distributed variables and linear relationships between them, you could use another function, shipley.test(), from the ggm library. (I would apologise for such an egotistical name but I did not create this library!) This function takes three arguments: amat, S, n. The first argument is a square Boolean matrix normally created using the DAG function, The second argument is the sample covariance matrix of the variables, obtained via the cov() function. The third argument is the sample size of the data set from which the covariance matrix was calculated. If you calculate your covariance matrix from the ranks of your data then you will get a result based on Spearman partial correlations.

[25] The model was actually developed using the exploratory methods of Chapter 8. This, too, should give us reason to question the model until independent data can be tested against it. At this point, all we can reasonably say is that the data are consistent with the model and so deserve further study.

[26] These methods are described in Chapters 4 to 7.

[27] One measured variable (the total number of seeds dispersed) was not provided to me, so I can't fit his original model.

[28] Analysis of covariance; a general linear model that blends ANOVA and regression.