1 Preliminaries

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

[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.

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,

[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.

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

[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 $t+\Delta t$, 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.

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

[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).

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

[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.

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

[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.

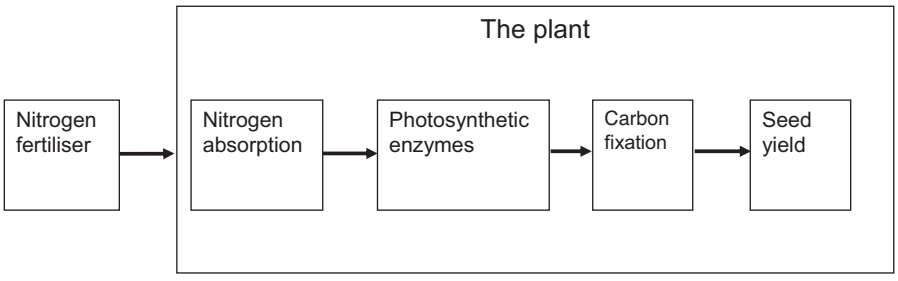

Figure 1.1 A hypothetical causal scenario that is not amenable to a randomised experiment

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).

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

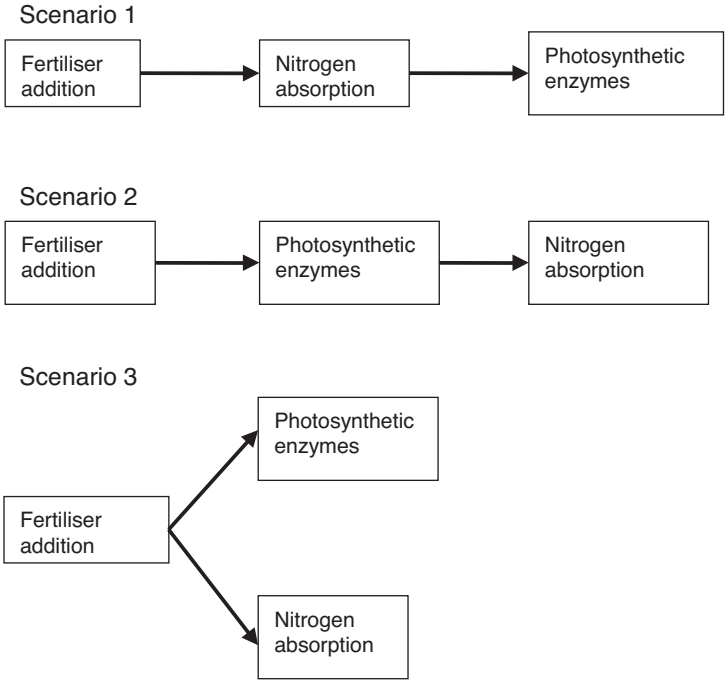

Figure 1.2 Three different causal scenarios that could generate an association between an increased nitrogen absorption and an increased enzyme concentration in the plant following the addition of fertiliser in a randomised experiment

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.

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.

[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.