Cause and Correlation in Biology

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

BILL SHIPLEY

Université de Sherbrooke,
Sherbrooke (Qc) Canada

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cup.org

© Cambridge University Press 2000

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2000
First paperback edition, with corrections 2002

Printed in the United Kingdom at the University Press, Cambridge

Typeface: Univers and Bembo 11/13pt.   System: QuarkXpress® [SE]

A catalogue record for this book is available from the British Library

Library of Congress cataloguing in publication data

Shipley, Bill, 1960–
Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference / Bill Shipley.
p. cm.
ISBN 0 521 79153 7 (hb)
1. Biometry. I. Title.

QH323.5.S477 2001
570′.1′5195 – dc21 00-037918

ISBN 0 521 79153 7 hardback

Contents

Preface xi

1 **Preliminaries** 1
- 1.1 The shadow's cause 1
- 1.2 Fisher's genius and the randomised experiment 7
- 1.3 The controlled experiment 14
- 1.4 Physical controls and observational controls 16

2 **From cause to correlation and back** 21
- 2.1 Translating from causal to statistical models 21
- 2.2 Directed graphs 25
- 2.3 Causal conditioning 28
- 2.4 d-separation 29
- 2.5 Probability distributions 32
- 2.6 Probabilistic independence 33
- 2.7 Markov condition 35
- 2.8 The translation from causal models to observational models 36
- 2.9 Counterintuitive consequences and limitations of d-separation: conditioning on a causal child 37
- 2.10 Counterintuitive consequences and limitations of d-separation: conditioning due to selection bias 41
- 2.11 Counterintuitive consequences and limitations of d-separation: feedback loops and cyclic causal graphs 42
- 2.12 Counterintuitive consequences and limitations of d-separation: imposed conservation relationships 43
- 2.13 Counterintuitive consequences and limitations of d-separation: unfaithfulness 45
- 2.14 Counterintuitive consequences and limitations of d-separation: context-sensitive independence 47
- 2.15 The logic of causal inference 48
- 2.16 Statistical control is not always the same as physical control 55
- 2.17 A taste of things to come 63

3 Sewall Wright, path analysis and d-separation 65
3.1 A bit of history 65
3.2 Why Wright's method of path analysis was ignored 66
3.3 d-sep tests 71
3.4 Independence of d-separation statements 72
3.5 Testing for probabilistic independence 74
3.6 Permutation tests of independence 79
3.7 Form-free regression 80
3.8 Conditional independence 83
3.9 Spearman partial correlations 88
3.10 Seed production in St Lucie's Cherry 90
3.11 Specific leaf area and leaf gas exchange 94

4 Path analysis and maximum likelihood 100
4.1 Testing path models using maximum likelihood 103
4.2 Decomposing effects in path diagrams 123
4.3 Multiple regression expressed as a path model 126
4.4 Maximum likelihood estimation of the gas-exchange model 130

5 Measurement error and latent variables 136
5.1 Measurement error and the inferential tests 138
5.2 Measurement error and the estimation of path coefficients 140
5.3 A measurement model 143
5.4 The nature of latent variables 152
5.5 Horn dimensions in Bighorn Sheep 157
5.6 Body size in Bighorn Sheep 158
5.7 Name calling 161

6 The structural equations model 162
6.1 Parameter identification 163
6.2 Structural underidentification with measurement models 164
6.3 Structural underidentification with structural models 171
6.4 Behaviour of the maximum likelihood chi-squared statistic with small sample sizes 173
6.5 Behaviour of the maximum likelihood chi-squared statistic with data that do not follow a multivariate normal distribution 179
6.6 Solutions for modelling non-normally distributed variables 185
6.7 Alternative measures of 'approximate' fit 188
6.8 Bentler's comparative fit index 192

6.9 Approximate fit measured by the root mean square error of approximation 193
6.10 An SEM analysis of the Bumpus House Sparrow data 195

7 Nested models and multilevel models 199
7.1 Nested models 200
7.2 Multigroup models 202
7.3 The dangers of hierarchically structured data 209
7.4 Multilevel SEM 221

8 Exploration, discovery and equivalence 237
8.1 Hypothesis generation 237
8.2 Exploring hypothesis space 238
8.3 The shadow's cause revisited 241
8.4 Obtaining the undirected dependency graph 243
8.5 The undirected dependency graph algorithm 246
8.6 Interpreting the undirected dependency graph 250
8.7 Orienting edges in the undirected dependency graph using unshielded colliders assuming an acyclic causal structure 254
8.8 Orientation algorithm using unshielded colliders 256
8.9 Orienting edges in the undirected dependency graph using definite discriminating paths 260
8.10 The Causal Inference algorithm 262
8.11 Equivalent models 264
8.12 Detecting latent variables 266
8.13 Vanishing Tetrad algorithm 271
8.14 Separating the message from the noise 272
8.15 The Causal Inference algorithm and sampling error 278
8.16 The Vanishing Tetrad algorithm and sampling variation 284
8.17 Empirical examples 287
8.18 Orienting edges in the undirected dependency graph without assuming an acyclic causal structure 294
8.19 The Cyclic Causal Discovery algorithm 299
8.20 In conclusion . . . 304

Appendix 305
References 308
Index 316

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 sees are not actors, nor even puppets, but rather 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. Shadows, however, can be ambiguous. In order to infer 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, sharpen their contours, and 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 to the underlying causal processes.

It is uncomfortably true, although rarely admitted in statistics texts, that many important areas of science are stubbornly impervious to experimental designs based on randomisation of treatments to experimental units. Historically, the response to this embarrassing problem has been to either 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, for instance structural equations modelling (SEM), are well known to researchers in other fields, although largely unknown to biologists. Other methods, for instance 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 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. 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 publicly claimed that inferring causes without randomisation and experimental control is probably impossible and, if possible, is not to be recommended (Shipley and Peters 1990). I had 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[1] and yet 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 Shipley and Peters (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 does not mean that such a method can't exist. In my defence, 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 learnt by almost every student who has ever taken an elementary course in statistics: *correlation does not imply causation*.

---

[1] 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.

In fact, with few exceptions[2], 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 sees a solid round shadow on the screen they know that some three-dimensional object has cast it, although they may not know whether 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, which is the effect, or even if both are common effects of some third, unmeasured variable.

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 fixed unambiguously. This point is developed more precisely in Chapter 2 but is so central to this book that it deserves repeating: 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 observe in observational data. The causal processes generating our observed data impose constraints on the patterns of correlation that such data display.

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 because 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.

---

[2] It could be argued that variables that covary because they are time-ordered have no causal basis. For instance, Monday unfortunately always follows Sunday and day always follows night. However, the first is simply a naming convention and there is a causal basis for the second: the earth's rotation about its axis in conjunction with its rotation around the sun. A more convincing example would be the correlation between the sizes of unrelated children, as they age, who are born at the same time.

Another reason why some prefer to speak of associations rather than causes is perhaps because causation is seen as a metaphysical notion that is best left to philosophers. In fact, even philosophers of science can't 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. Where 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, although, 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[3].

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) gave 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 define 'probability' cannot agree on one does not prevent biologists from effectively using probabilities, significance levels, confidence intervals, and the other paraphernalia of modern statistics[4]. In fact, insisting on such an agreement would mean that modern statistics could not even have begun.

---

[3] 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.

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 can't 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 a reliable means of transport.

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 develop 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 can be intuitively understood to mean that events are caused only by their proximate causes. Thus, if event A causes event C

---

[4] 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.

only through its effect of an intermediate event B ($A \rightarrow B \rightarrow C$), 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[5]. Finally, these relationships must be *asymmetric*: if A is a cause of B, then B cannot simultaneously be a cause of A[6]. 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[7] (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.

---

[5] 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'.

[6] 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.

[7] 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.

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 comfortably speak about causes? What is it about inferring causality from non-experimental data that make them squirm in their chairs? I will distinguish between two basic types of experiment: controlled and randomised. Although the controlled experiment takes historical precedence, the randomised experiment takes precedence in the strength of its causal inferences.

Fisher[8] described the principles of the randomised experiment in his classic *The design of experiments* (Fisher 1926). Since he developed many of his statistical methods in the context of agronomy, let's consider a typical randomised experiment designed to determine whether 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 (50 cm × 50 cm) 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 6. Those plots receiving 20 kg/hectare of fertiliser produce 80 g of seed with a standard error of 6. 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 of 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[9]. Yet Fisher and many other statisticians[10] since (Kempthorpe 1979; Kendall and Stuart 1983) claim further 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?

---

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

Given that two variables (X and Y) are associated, there can be only three elementary, but not mutually exclusive, causal explanations: 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 to be added to any given plot: the number that the experimenter saw written on the piece of paper attributed to that plot. Second, the fertiliser was added before the wheat plants began to produce seed[11]. 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.

---

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

[10] '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).

[11] 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. Like pornography, we have difficulty defining it but we all seem to know it when we see it.

Let's retrace the logical steps. We began by asserting that, if there was a causal relationship between fertiliser addition and seed yield, then there would also be a systematic relationship between these two variables in our data: *causation implies correlation*. When we observe a systematic relationship that can't 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 of fertiliser per hectare 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, we can only advance an argument that is beyond reasonable doubt, not a logical certainty.

The key role played by the process of randomisation seems to be to ensure, 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[12]. 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 is a large number of potential causes of seed yield, then 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.

---

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

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):

> 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 does emphasise that even the randomised experiment does not provide any automatic assurance of causal inference, free from 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[13] (see, for example, Feiblman 1972, page 149). R. A. Fisher (1970) explicitly rejected this as an inferior method, pointing out that it is logically impossible to know whether 'all other variables' have been accounted for. This is not to say that Fisher did not advocate physically controlling for other causes in addition to randomisation. In fact, he explicitly recommended that the researcher do this whenever possible. For instance, in discussing the comparison of plant yields of different varieties, he advised that they be planted in soil 'that appears to be uniform'. In the context of pot experiments he recommended that the soil be thor-

---

[13] Clearly, this cannot be literally true. Consider a case in which the causal process is: $A \rightarrow B \rightarrow C$ 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, although 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.

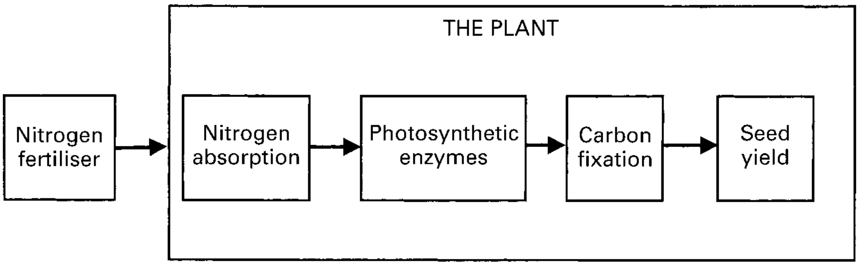

Figure 1.1. An hypothetical causal scenario that is not amenable to a randomised experiment.

oughly mixed before putting it in the pots, that the watering be equalised, that they receive the same amount of light and so on. The strength of the randomised experiment is 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. Yet, a large proportion of biological studies involves 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 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.

While 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 (for instance 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[14]. When such random assignments can't be done then one can't 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 can't 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 whether increased carbon dioxide (CO_2) con-

---

[14] 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 can't randomly add different numbers of functioning chloroplasts to a leaf.

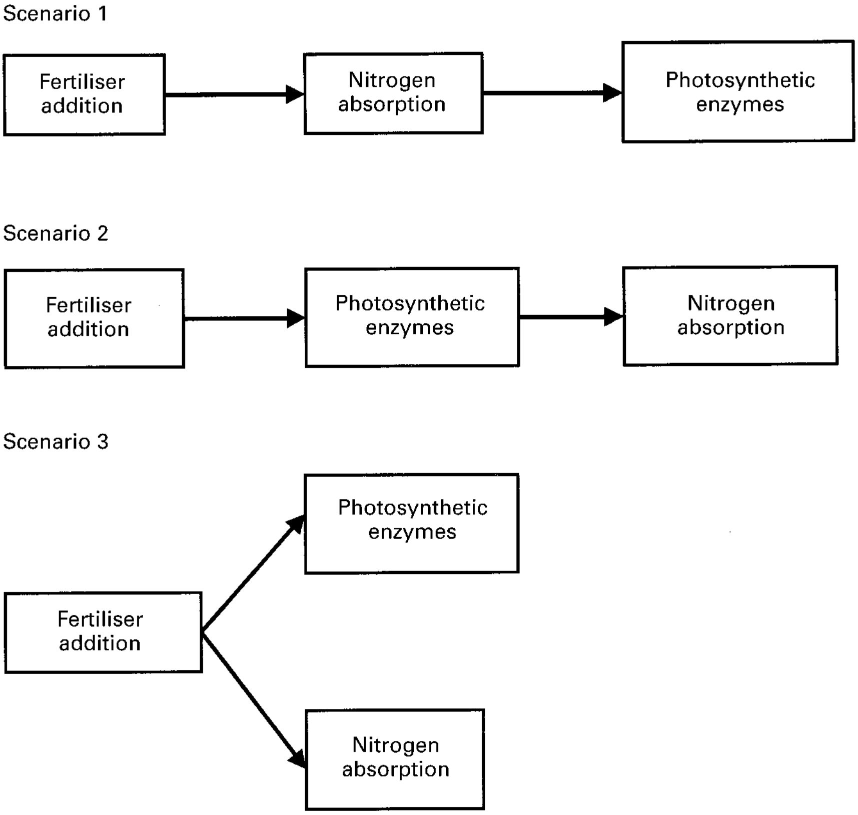
Scenario 1
Fertiliser addition
Nitrogen absorption
Photosynthetic enzymes
Scenario 2
Fertiliser addition
Photosynthetic enzymes
Nitrogen absorption
Scenario 3
Photosynthetic enzymes
Fertiliser addition
Nitrogen absorption

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

centrations 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 sometimes can't be done in practice, owing 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 seem 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[15]. 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 neighboring 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 coloring 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 coloring 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 coloring of blood in the renal vein.

Our knowledge of human physiology has progressed far from the experiments of Claude 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 to much of biological research. The controlled experi-

---

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

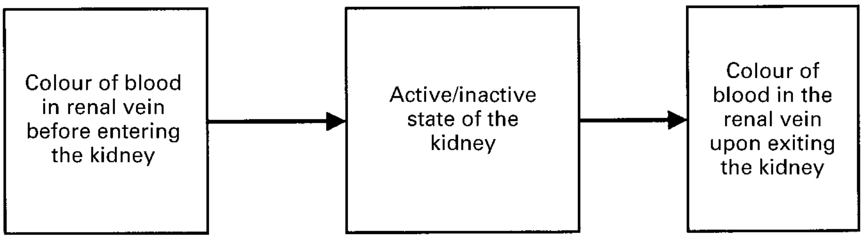
Colour of blood in renal vein before entering the kidney
Active/inactive state of the kidney
Colour of blood in the renal vein upon exiting the kidney

Figure 1.3. The hypothetical causal explanation invoked by Claude Bernard.

ment 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 Claude Bernard, the hypothetical causal structure could be conceptualised as shown in Figure 1.3.

The key notion in Bernard's experiment was the realisation that, if his causal explanation were true, then 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 coloring of blood in the renal vein.' Since he explicitly stated earlier in the quote that he was inquiring 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 on the basis of 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 autumn. However, this species has a strong sexual dimorphism, males being up to 60% 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.

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

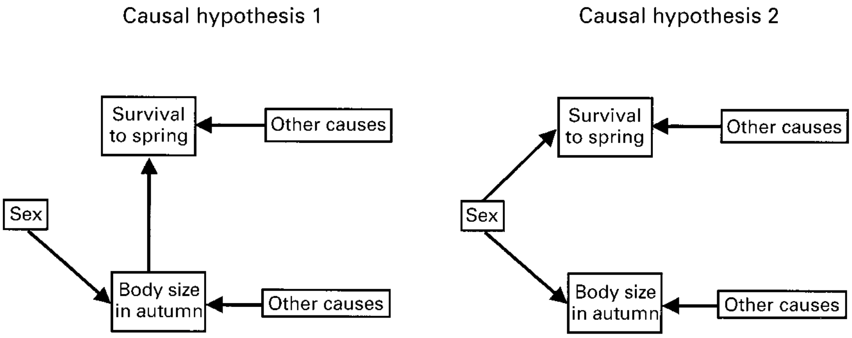

Figure 1.4. Two alternative causal explanations for the relationship between sex, body size of Bighorn Sheep in the autumn and the probability of survival until the spring.

the sex ratio of the population. Thus, if the sex ratio could be held constant, then 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 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[16]. 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 certainly conceivable 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 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

---

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

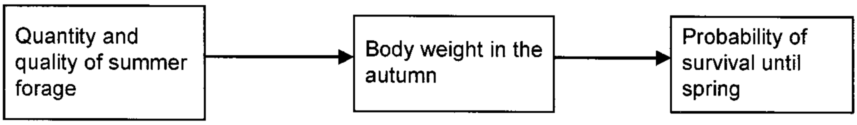
Quantity and quality of summer forage
Body weight in the autumn
Probability of survival until spring

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

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 sheep 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 then increases their chances of survival during the winter (Figure 1.5).

The logic of the controlled experiment requires that we be able to compare the relationship between forage quality and winter survival after physically preventing body weight from changing, which we can't do[17]. 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 for 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).

The formula for a linear regression is: $Y_i = \alpha + \beta X_i + \mathrm{N}(0,\sigma)$. Here, the notation '$\mathrm{N}(0,\sigma)$' 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 '$\mathrm{E}(Y|X)$' represent the expected value of Y given X, then we can write:

---

[17] 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 whether 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 whether body size in the autumn is a cause of increased survival during the winter.

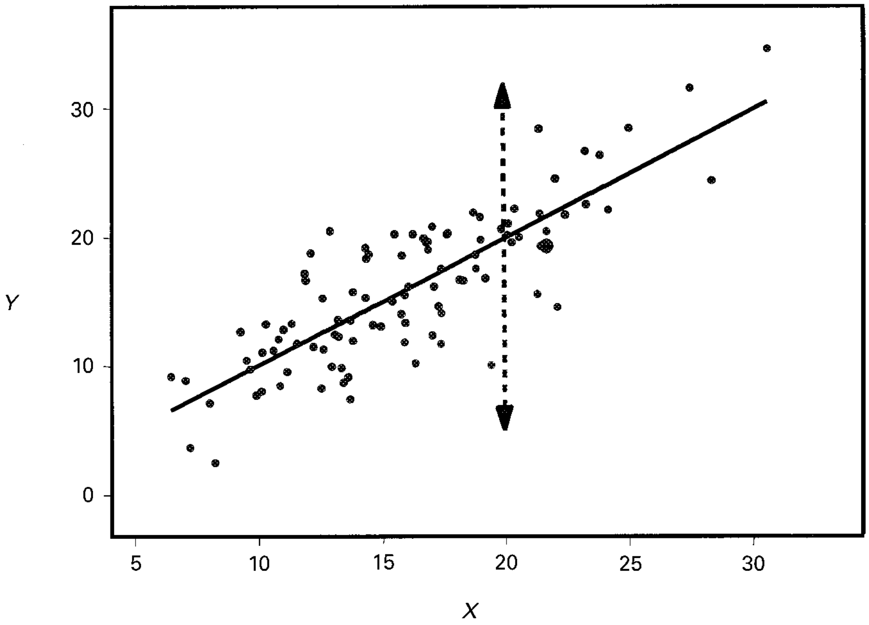

Figure 1.6. A simple bivariate regression. The solid line shows the expected value of Y_i given the value of X_i ($E[Y_i|X_i]$). The dotted line shows the possible values of Y_i that are independent of X_i (the residuals).

$$E(Y|X_i) = \alpha + \beta X_i$$

$$Y_i = E(Y|X_i) + N(0,\sigma)$$

$$Y_i - E(Y|X_i)) = N(0,\sigma).$$

Thus, if we subtract the expected value of each Y, given X, from the value itself, then 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 for 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 its given body weight, and then see whether 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 patterns of association different from those that would occur when one is 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.