Preamble

This book is about the planning and analysis of a special kind of investigation: a *case-control study*. We use this term to cover a number of different designs. In the simplest form individuals with an outcome of interest, possibly rare, are observed and information about past experience is obtained. In addition corresponding data are obtained on suitable controls in the hope of explaining what influences the outcome. In this book we are largely concerned with binary outcomes, for example indicating disease diagnosis or death. Such studies are reasonably called *retrospective* as contrasted with *prospective* studies, in which one records explanatory features and then waits to see what outcome arises. In retrospective studies we are studying the causes of effects and in prospective studies we are studying the effects of causes. We also discuss some extensions of case-control studies to incorporate temporality, which may be more appropriately viewed as a form of prospective study. The key aspect of all these designs is that they involve a sample of the underlying population that motivates the study, in which individuals with certain outcomes are strongly over-represented.

While we shall concentrate on the many special issues raised by such studies, we begin with a brief survey of the general themes of statistical design and analysis. We use a terminology deriving in part from epidemiological applications although the ideas are of much broader relevance.

We start the general discussion by considering a population of study individuals, patients, say, assumed to be statistically independent. The primary object is to understand the effect of exposures (or treatments or conditions) on an outcome or response. Exposures are represented by a random variable X and the outcome by a random variable Y, where typically X and sometimes Y are vectors. Our interest is in the effect of X on Y. This relationship can be represented in a diagram as in Figure 1. The arrow points from one variable to another, which is in some sense its outcome; that is, the arrow indicates that X has some effect or influence on Y. The direction of the arrow is usually that of time. Such diagrams are described as path diagrams

Figure 1 The effect of an exposure X on an outcome Y. The arrow represents statistical dependence directed from X to Y.

or, in the special case where all the variables included are one dimensional, as *directed acyclic graphs* or DAGS. We use path diagrams to help illustrate some issues of study design and analysis.

We define a third class of variables, referred to as *intrinsic variables* and represented by the vector random variable W. Depending on the study setting these variables may affect X or Y or both, an aspect which we discuss further below. The inter-relationships between W, X and Y in a given study setting affect how we estimate the effect of X on Y; that is, the possible involvement of W should be considered. In this specification both X and W are explanatory variables. The distinction between them is one of subject matter and is context dependent and not to be settled by a formal statistical test. For a variable to be considered as an exposure it has to be relevant, even if not realizable, to ask: how would the outcome of an individual have changed had their exposure been different from what it is, *other things being equal*? By contrast, W represents properties of individuals that are immutable in the context in question. We take 'other things being equal' to mean that the intrinsic variable, W, is fixed when one is studying the possible effect on Y of changing the exposure X. Because the variables W are intrinsic, they typically refer to a time point prior to the exposure X.

There are four broad ways in which the systems described above may be investigated:

- by **randomized experiment**;
- by **prospective observational study**, that is, cohort study;
- by **retrospective observational study**, that is, case-control study;
- by **cross-sectional observational study**.

We describe each type of study in turn. In a **randomized experiment** the level of the exposure X for each study individual is assigned by the investigator using a randomizing device, which ensures that each individual is equally likely to receive each of a set of exposure levels. Examples of exposures are medical treatments received by a patient and fertilizer treatments in an agricultural trial. The outcome Y is recorded after a suitable time. The relationship between X, W and Y in a simple randomized experiment is illustrated in Figure 2, where R denotes the randomization process.

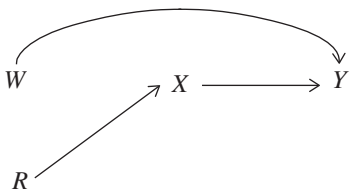

Figure 2 Simple randomized experiment. Randomization R disconnects X and all intrinsic variables W.

In the formulation represented in Figure 2 the exposure X is determined entirely by the randomization process and therefore the exposure for any specific individual is independent of all the intrinsic variables W. As a result, ignoring W produces no systematic distortion when estimating the effect of X on Y.

If W has an effect on Y and the latter is a continuous variable analysed by normal-theory methods then a component of the variance of Y may be in effect eliminated by regression on W and the precision of the resulting assessment of the effect of X thereby enhanced. With a binary Y, the situation with which we are mainly concerned, things are more complicated. In both cases the possibility of an X-by-W interaction may, however, be important.

In an observational study of a population the exposure X is determined by the individual, or as a result of their circumstances, as opposed to being set by the investigator through experimental design. Examples of such exposures are industrial or environmental hazards in the workplace, smoking by individuals and so on. For a patient in a clinical study, W may include gender and age at initial diagnosis or the age at entry into the study. In relatively simple cases X is one dimensional, whereas W is typically multidimensional.

The structure of the data in a **prospective observational study** may appear essentially the same as in a randomized experiment, but in the former there is the crucial distinction that the exposure, X, of each individual is outside the investigator's control. The intrinsic variables W may thus influence X as well as Y.

Suppose that the response variable Y is binary with one outcome having a very low frequency. An individual having this outcome, the outcome of particular interest, is known as a *case*. Then a prospective observational study may be very inefficient, in that large amounts of data may be collected on non-cases, or *controls*, when effectively the same precision for

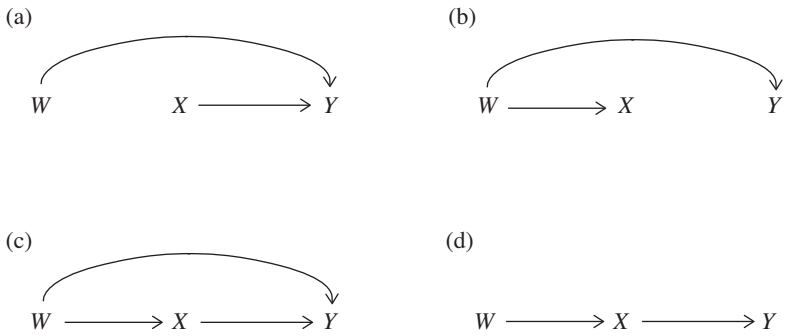

Figure 3 General specification: relationships between exposures X, intrinsic variables W and outcome Y in observational studies.

the case-versus-control comparison would be achieved with many fewer controls. This is one motivation for the third kind of study, the **retrospective observational study**. In this, cases are identified first and then control individuals selected and the variables X and W found retrospectively. The method of selecting controls is crucial and is discussed in more detail in the next chapter. In a retrospective study the object remains, as in a prospective study, to determine the conditional dependence of Y on X given W in the population, but, because of the special methods of selecting individuals used for a retrospective study, this relation is addressed indirectly.

Figure 3 illustrates four different types of interrelationship between X, W and Y which may arise in prospective and retrospective observational studies.

In Figure 3(a) W affects Y but not X. Hence W does not interfere in the effect of X on Y. For a continuous outcome, ignoring W causes no systematic distortion in the estimated effect of X on Y, though controlling for W in a regression adjustment may produce some improvement in precision as noted for randomized experiments. For a binary outcome, however, the situation is different. When the association between Y and X given W is described by the conditional odds ratio between Y and X given W, which is the commonly used effect measure used in case-control studies, this is not the same as the marginal odds ratio between Y and X even if W is not directly connected to X. This is a feature of odds ratios called non-collapsibility and is discussed further in later chapters.

In Figure 3(b) W affects both Y and X, but there is no effect of X on Y given W, indicated by the absence of an arrow from X to Y. An analysis

that ignores W would tend to find a spurious non-zero effect of X on Y because of the common effect of W.

Figure 3(c) is the same as Figure 3(b) except for the inclusion of an arrow from X to Y. Failure to account for W in this situation would result in a biased estimate of the effect of X on Y given W. The systematic distortion when there are two paths between X and Y as in Figure 3(c) is called *confounding*. By conditioning on W, which is necessary to assess the effect of changing X while leaving the past unchanged, we close the extra path from X to Y via W and obtain an unbiased estimate of the effect of X on Y.

In Figure 3(d) W affects X but not Y. That is, Y is independent of W given X. Then aspects of the conditional relation between Y and X given W are correctly estimated by ignoring W; that is, it is not necessary to account for W in the analysis, except possibly to examine X-by-W interactions.

Another approach to investigation is by a **cross-sectional observational study**, in which data are collected that refer only to the status of individuals at the instant of observation. This may give useful information on correlations but, in such a study, if two variables are correlated then it is in principle impossible to say from the study alone which variable is the response and which is explanatory; any conclusion of this sort must rest on external information or assumptions. We do not consider cross-sectional studies further here.

In observational studies, an important role of the intrinsic variables W is to adjust for the dependences of X and Y on W, that is, to remove systematic error or *confounding*. This is in contrast with randomized experiments, where the roles of W are precision improvement and interaction detection, always assuming that the randomization has been effective. In some contexts W might be extended to include variables that are not intrinsic features; these variables are defined conceptually prior to the exposure, and, unless accounted for in some way, could distort the estimated association between exposure and outcome. For example, consider a study of the association between a patient treatment and a medical outcome. It might be important to include in W information about other patient medication, use of which may be associated with whether the patient received the treatment of interest and also with the outcome.

A major cause for concern in observational studies is that some components of W may be unobserved. Let W_O and W_U denote the observed and unobserved components respectively. Figure 4 illustrates a situation where both W_O and W_U have arrows to X and Y, that is they are both confounders of the effect of X on Y. Confounding by the observed intrinsic

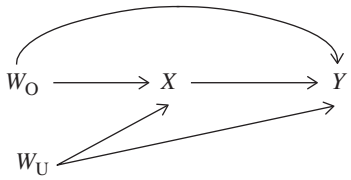

Figure 4 General specification: confounding by observed variables W_O and unmeasured variables W_U.

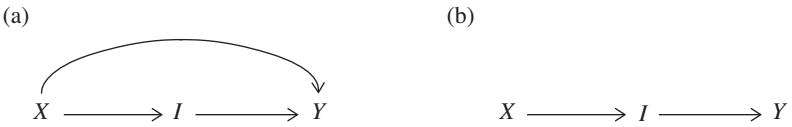

Figure 5 General specification: the presence of an intermediate variable I.

variables W_O can be controlled by conditioning on the W_O. However, because W_U is unobserved its effect cannot be controlled for and the possible role of an unobserved W_U often limits the security of the interpretation. If the arrow from W_U to Y is missing then there is no confounding by the unmeasured variables and conditioning on W_U is not necessary to obtain unbiased estimates of the effect of X on Y. If the arrow from W_U to X is missing but there remains an arrow from W_U to Y then there is no confounding but we may still face a systematic change in the observed association if W_U is ignored in some special situations, notably when the association between X and Y is measured using an odds ratio. This relates to the discussion around Figure 3(a).

We have focused on the role of the intrinsic variables W in the effect of X on Y. We now consider whether the effect of X on Y may possibly act though an intermediate variable, I say (Figure 5). First, in the primary analysis of the effect of X on Y the intermediate response I is ignored, that is, marginalized. In a subsidiary analysis we may consider the mediating effect of I, in particular whether any effect of X on Y is explained largely or even entirely through the effect of X on I, in which case Y is conditionally independent of X given I (Figure 5(b)). In other situations I is itself of interest and is analysed as a response on its own, ignoring Y.

In summary, it is crucial in considering the relation between exposure and outcome to include appropriate conditioning variables W, and to exclude inappropriate ones, and to exclude intermediate variables I when studying the total effect of X on Y.

The ideas sketched here can be developed in many directions, for example to mixtures of study types and to complications arising when there are different exposure variables at different time points for the same individual.

Notes

See Wright (1921) for an early discussion of correlation and causation. Greenland *et al.* (1999) give an introductory account of directed acyclic graphs with an emphasis on epidemiological applications. For discussions of the use of directed acyclic graphs in case-control studies, see Hernán *et al.* (2004), Didelez *et al.* (2010) and Mansournia *et al.* (2013).

1

Introduction to case-control studies

- A case-control study is a retrospective observational study and is an alternative to a prospective observational study. Cases are identified in an underlying population and a comparable control group is sampled.
- In the standard design exposure information is obtained retrospectively, though this is not necessarily the case if the case-control sample is nested within a prospective cohort.
- Prospective studies are not cost effective for rare outcomes. By contrast, in a case-control study the ratio of cases and controls is higher than in the underlying population in order to make more efficient use of resources.
- There are two main types of case-control design; matched and unmatched.
- The odds ratio is the most commonly used measure of association between exposure and outcome in a case-control study.
- Important extensions to the standard case-control design include the explicit incorporation of time into the choice of controls and into the analysis.

1.1 Defining a case-control study

Consider a population of interest, for example the general population of the UK, perhaps restricted by gender or age group. We may call a representation of the process by which *exposures* X and *outcomes* Y occur in the presence of intrinsic features W the *population model*. As noted in the Preamble, such a system may be investigated prospectively or retrospectively; see Figure 1.1. In a prospective or cohort study a suitable sample of individuals is chosen to represent the population of interest, values of (W, X) are determined and the individuals are followed through time until the outcome Y can be observed. By contrast, in a retrospective case-control study, the primary subject of this book, we start with individuals observed to have a specific outcome, say $Y = 1$, whom we call *cases*, and then choose a suitable number of *controls*, often one control for each case. For the

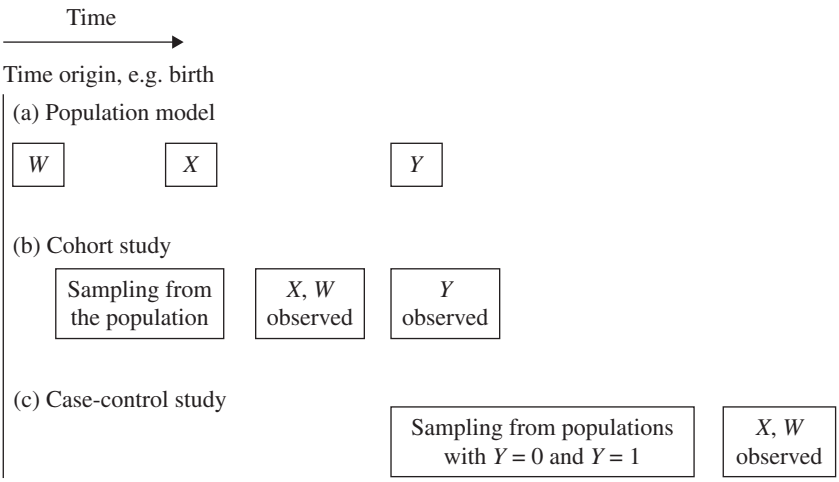

Figure 1.1 (a) Population model with intrinsic features W present; (b) cohort study (prospective); (c) case-control study (retrospective).

controls, members of the population at risk who are not cases, we define Y to be zero. Values of (W, X) are then determined retrospectively on the chosen individuals. The cases are chosen to represent those occurring in the population of interest, and the controls are chosen to represent the part of the population of interest with $Y = 0$.

The essence of a case-control study is that we start with the outcome and look back to find the exposures, that is, the explanatory features of concern. Another core characteristic is that the ratio of cases and controls is not the same as in the population. Indeed typically in the population that ratio is very small because the outcome is rare, whereas in the case-control study it can be as high as one to one. Case-control studies can, however, also be used in situations where the outcome is common rather than rare.

Different procedures for sampling cases and controls lead to several different forms of case-control design. The 'standard' case-control design is as follows and is illustrated in Figure 1.1(c). The cases consist in principle of all individuals in a specific population who have experienced the outcome in question within a specified, usually short, period of time, be that calendar time or age. More generally one might take a random sample of such cases. The control group is sampled from those individuals in the same population who were eligible to experience the outcome during the specified time

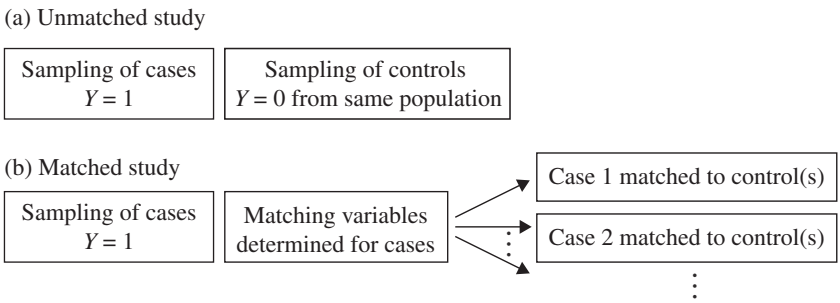

Figure 1.2 (a) An unmatched case-control study; (b) a matched case-control study, in which the cases are matched to one or more controls.

period but did not do so. For both cases and controls, exposure measures and intrinsic variables are then determined.

Within this standard framework there are two main forms of case-control study, unmatched studies and matched studies; see Figure 1.2. In the first, an *unmatched case-control study*, a shared control group for all cases is selected essentially at random, or, more often, at random given a set of intrinsic variables, perhaps such that the distribution of certain intrinsic variables is similar among the case group and the control group. In the second form, a *matched case-control study*, controls are selected case by case in such a way that they are constrained to match individual cases in certain specified respects, that is, so that to each case is attached one or more controls. Matching on variables that confound the effect of exposure on outcome is a way of conditioning on the confounders. Methods for dealing with confounding in both unmatched and matched studies are discussed in more detail in Section 1.3.

The path diagrams used in the Preamble refer to the study population underlying a case-control study, but they can be extended to incorporate case-control sampling. We define on the underlying study population a binary indicator variable D taking the value 1 for individuals in the case-control sample and taking the value 0 otherwise. Figure 1.3(a) extends previous diagrams to introduce D; we include the possibility of confounding by intrinsic variables W. The arrow from Y to D arises because, in the study population, individuals with $Y = 1$ are much more likely to be sampled to the case-control study than individuals with $Y = 0$. The case-control study corresponds to those with $D = 1$ and hence any analysis is conditional on $D = 1$; the conditioning is indicated by the box around