CHAPTER SIX

Simpson's Paradox, Confounding, and Collapsibility

He who confronts the paradoxical
exposes himself to reality.
Friedrick Durrenmatt (1962)

Preface

Confounding represents one of the most fundamental impediments to the elucidation of causal inferences from empirical data. As a result, the consideration of confounding underlies much of what has been written or said in areas that critically rely on causal inferences; this includes epidemiology, econometrics, biostatistics, and the social sciences. Yet, apart from the standard analysis of randomized experiments, the topic is given little or no discussion in most statistics texts. The reason for this is simple: confounding is a causal concept and hence cannot be expressed in standard statistical models. When formal statistical analysis is attempted, it often leads to confusions or complexities that make the topic extremely hard for the nonexpert to comprehend, let alone master.

One of my main objectives in writing this book is to see these confusions resolved – to see problems involving the control of confounding reduced to simple mathematical routines. The mathematical techniques introduced in Chapter 3 have indeed culminated in simple graphical routines for detecting the presence of confounding and for identifying variables that should be controlled in order to obtain unconfounded effect estimates. In this chapter, we address the difficulties encountered when we attempt to define and control confounding by using statistical criteria.

We start by analyzing the interesting history of Simpson's paradox (Section 6.1) and use it as a magnifying glass to examine the difficulties that generations of statisticians have had in their attempts to capture causal concepts in the language of statistics. In Sections 6.2 and 6.3, we examine the feasibility of replacing the causal definition of confounding with statistical criteria that are based solely on frequency data and measurable statistical associations. We will show that, although such replacement is generally not feasible (Section 6.3), a certain kind of nonconfounding conditions, called *stable*, can be given statistical or semistatistical characterization (Section 6.4). This characterization leads to operational tests, similar to collapsibility tests, that can alert investigators to the existence of either instability or bias in a given effect estimate (Section 6.4.3). Finally, Section 6.5 clarifies distinctions between collapsibility and no-confounding, confounders and confounding, and between the structural and exchangeability approaches to representing problems of confounding.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

6.1 SIMPSON'S PARADOX: AN ANATOMY

The reversal effect known as Simpson's paradox has been briefly discussed twice in this book: first in connection with the covariate selection problem (Section 3.3) and then in connection with the definition of direct effects (Section 4.5.3). In this section we analyze the reasons why the reversal effect has been (and still is) considered paradoxical and why its resolution has been so late in coming.

6.1.1 A Tale of a Non-Paradox

Simpson's paradox (Simpson 1951; Blyth 1972), first encountered by Pearson in 1899 (Aldrich 1995), refers to the phenomenon whereby an event C increases the probability of E in a given population p and, at the same time, decreases the probability of E in every subpopulation of p. In other words, if F and $\neg F$ are two complementary properties describing two subpopulations, we might well encounter the inequalities

$$P(E \mid C) > P(E \mid \neg C), \tag{6.1}$$

$$P(E \mid C, F) < P(E \mid \neg C, F), \tag{6.2}$$

$$P(E \mid C, \neg F) < P(E \mid \neg C, \neg F). \tag{6.3}$$

Although such order reversal might not surprise students of probability, it is paradoxical when given causal interpretation. For example, if we associate C (connoting *cause*) with taking a certain drug, E (connoting *effect*) with recovery, and F with being a female, then – under the causal interpretation of (6.2)–(6.3) – the drug seems to be harmful to both males and females yet beneficial to the population as a whole (equation (6.1)). Intuition deems such a result impossible, and correctly so.

The tables in Figure 6.1 represent Simpson's reversal numerically. We see that, overall, the recovery rate for patients receiving the drug (C) at 50% exceeds that of the control ($\neg C$) at 40%, and so the drug treatment is apparently to be preferred. However, when we inspect the separate tables for males and females, the recovery rate for the untreated patients is 10% higher than that for the treated ones, for males and females both.

The explanation for Simpson's paradox should be clear to readers of this book, since we have taken great care in distinguishing *seeing* from *doing*. The conditioning operator in probability calculus stands for the evidential conditional "given that we see," whereas the $do(\cdot)$ operator was devised to represent the causal conditional "given that we do." Accordingly, the inequality

$$P(E \mid C) > P(E \mid \neg C)$$

is not a statement about C having a positive effect on E, properly written

$$P(E \mid do(C)) > P(E \mid do(\neg C)),$$

but rather about C being positive *evidence* for E, which may be due to spurious confounding factors that cause both C and E. In our example, the drug appears beneficial

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

	Combined	E	$\neg E$		Recovery Rate
(a)	Drug (C)	20	20	40	50%
	No drug ($\neg C$)	16	24	40	40%
		36	44	80	
	Males	E	$\neg E$		Recovery Rate
(b)	Drug (C)	18	12	30	60%
	No drug ($\neg C$)	7	3	10	70%
		25	15	40	
	Females	E	$\neg E$		Recovery Rate
(c)	Drug (C)	2	8	10	20%
	No drug ($\neg C$)	9	21	30	30%
		11	29	40	

Figure 6.1 Recovery rates under treatment (C) and control ($\neg C$) for males, females, and combined.

overall because the males, who recover (regardless of the drug) more often than the females, are also more likely than the females to use the drug. Indeed, finding a drug-using patient (C) of unknown gender, we would do well inferring that the patient is more likely to be a male and hence more likely to recover, in perfect harmony with (6.1)–(6.3).

The standard method for dealing with potential confounders of this kind is to "hold them fixed,"[1] namely, to condition the probabilities on any factor that might cause both C and E. In our example, if being a male ($\neg F$) is perceived to be a cause for both recovery (E) and drug usage (C), then the effect of the drug needs to be evaluated separately for men and women (as in (6.2)–(6.3)) and then averaged accordingly. Thus, assuming F is the only confounding factor, (6.2)–(6.3) properly represent the efficacy of the drug in the respective populations, while (6.1) represents merely its evidential weight in the absence of gender information, and the paradox dissolves.

6.1.2 A Tale of Statistical Agony

Thus far, we have described the paradox as it is understood, or should be understood, by modern students of causality (see, e.g., Cartwright 1983;[2] Holland and Rubin 1983; Greenland and Robins 1986; Pearl 1993b; Spirtes et al. 1993; Meek and Glymour 1994). Most

[1] The phrases "hold F fixed" and "control for F," used by both philosophers (e.g., Eells 1991) and statisticians (e.g., Pratt and Schlaifer 1988), connote external interventions and may therefore be misleading. In statistical analysis, all one can do is *simulate* "holding F fixed" by considering cases with equal values of F – that is, "conditioning" on F and $\neg F$ – an operation that I will call "adjusting for F."

[2] Cartwright states, though, that the third factor F should be "held fixed" if and only if F is causally relevant to E (p. 37); the correct (back-door) criterion is somewhat more involved (see Definition 3.3.1).

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

statisticians, however, are reluctant to entertain the idea that Simpson's paradox emerges from causal considerations. The general attitude is as follows: The reversal is real and disturbing, because it actually shows up in the numbers and may actually mislead statisticians into incorrect conclusions. If something is real, then it cannot be causal, because causality is a mental construct that is not well defined. Thus, the paradox must be a statistical phenomenon that can be detected, understood, and avoided using the tools of statistical analysis. *The Encyclopedia of Statistical Sciences*, for example, warns us sternly of the dangers lurking from Simpson's paradox with no mention of the words "cause" or "causality" (Agresti 1983). *The Encyclopedia of Biostatistics* (Dong 1998) and *The Cambridge Dictionary of Statistics in Medical Sciences* (Everitt 1995) uphold the same conception.

I know of only two articles in the statistical literature that explicitly attribute the peculiarity of Simpson's reversal to causal interpretations. The first is Pearson et al. (1899), where the discovery of the phenomenon[3] is enunciated in these terms:

> To those who persist on looking upon all correlation as cause and effect, the fact that correlation can be produced between two quite uncorrelated characters *A* and *B* by taking an artificial mixture of the two closely allied races, must come as rather a shock. (p. 278)

Influenced by Pearson's life-long campaign, statisticians have refrained from causal talk whenever possible and, for over half a century, the reversal phenomenon has been treated as a curious mathematical property of 2×2 tables, stripped of its causal origin. Finally, Lindley and Novick (1981) analyzed the problem from a new angle, and made the second published connection to causality:

> In the last paragraph the concept of a "cause" has been introduced. One possibility would be to use the language of causation, rather than that of exchangeability or identification of populations. We have not chosen to do this; nor to discuss causation, because the concept, although widely used, does not seem to be well-defined. (p. 51)

What is amazing about the history of Simpson's reversal is that, from Pearson et al. to Lindley and Novick, none of the many authors who wrote on the subject dared ask why the phenomenon should warrant our attention and why it evokes surprise. After all, seeing probabilities change magnitude upon conditionalization is commonplace, and seeing such changes turn into sign reversal (by taking differences and mixtures of those probabilities) is not uncommon either. Thus, if it were not for some misguided yet persistent illusion, what is so shocking about inequalities reversing direction?

Pearson understood that the shock originates with distorted causal interpretations, which he set out to correct through the prisms of statistical correlations and contingency tables (see the Epilogue following Chapter 10). His disciples took him rather seriously, and some even asserted that causation is none but a species of correlation (Niles 1922). In so denying any attention to causal intuition, researchers often had no choice but to attribute Simpson's reversal to some evil feature of the data, one that ought to be avoided

[3] Pearson et al. (1899) and Yule (1903) reported a weaker version of the paradox in which (6.2)–(6.3) are satisfied with equality. The reversal was discovered later by Cohen and Nagel (1934, p. 449).

by scrupulous researchers. Dozens of papers have been written since the 1950s on the statistical aspects of Simpson's reversal; some dealt with the magnitude of the effect (Blyth 1972; Zidek 1984), some established conditions for its disappearance (Bishop et al. 1975; Whittemore 1978; Good and Mittal 1987; Wermuth 1987), and some even proposed remedies as drastic as replacing $P(E \mid C)$ with $P(C \mid E)$ as a measure of treatment efficacy (Barigelli and Scozzafava 1984) – the reversal had to be avoided at all cost.

A typical treatment of the topic can be found in the influential book of Bishop, Fienberg, and Holland (1975). Bishop et al. (1975, pp. 41–2) presented an example whereby an apparent association between amount of prenatal care and infant survival disappears when the data are considered separately for each clinic participating in the study. They concluded: "If we were to look only at this [the combined] table we would erroneously conclude that survival *was related* [my italics] to the amount of care received." Ironically, survival *was* in fact *related* to the amount of care received in the study considered. What Bishop et al. meant to say is that, looking uncritically at the combined table, we would erroneously conclude that survival was *causally* related to the amount of care received. However, since causal vocabulary had to be avoided in the 1970s, researchers like Bishop et al. were forced to use statistical surrogates such as "related" or "associated" and so naturally fell victim to the limitations of the language; statistical surrogates could not express the causal relationships that researchers meant to convey.

Simpson's paradox helps us to appreciate both the agony and the achievement of this generation of statisticians. Driven by healthy causal intuition, yet culturally forbidden from admitting it and mathematically disabled from expressing it, they managed nevertheless to extract meaning from dry tables and to make statistical methods the standard in the empirical sciences. But the spice of Simpson's paradox turned out to be nonstatistical after all.

6.1.3 Causality versus Exchangeability

Lindley and Novick (1981) were the first to demonstrate the nonstatistical character of Simpson's paradox – that there is no statistical criterion that would warn the investigator against drawing the wrong conclusions or would indicate which table represents the correct answer.

In the tradition of Bayesian decision theory, they first shifted attention to the practical side of the phenomenon and boldly asked: A new patient comes in; do we use the drug or do we not? Equivalently: Which table do we consult, the combined or the gender-specific? "The apparent answer is," confesses Novick (1983, p. 45), "that when we know that the gender of the patient is male or when we know that it is female we do not use the treatment, but if the gender is unknown we should use the treatment! Obviously that conclusion is ridiculous." Lindley and Novick then go through lengthy informal discussion, concluding (as we did in Section 6.1.1) that we should consult the gender-specific tables and not use the drug.

The next step was to ask whether some additional statistical information could in general point us to the right table. This question Lindley and Novick answered in the negative by showing that, with the very same data, we sometimes should decide the opposite and

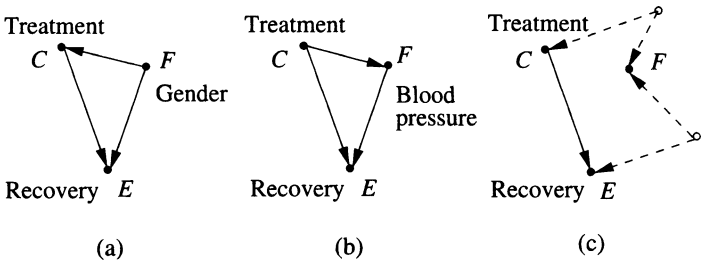

Figure 6.2 Three causal models capable of generating the data in Figure 6.1. Model (a) dictates use of the gender-specific tables, whereas (b) and (c) dictate use of the combined table.

consult the combined table. They asked: Suppose we keep the same numbers and merely change the story behind the data, imagining that F stands for some property that is affected by C – say, low blood pressure, as shown in Figure 6.2(b).[4] By inspecting the diagram in Figure 6.2(b), the reader should immediately conclude that the combined table represents the answer we want; we should not condition on F because it resides on the very causal pathway that we wish to evaluate. (Equivalently, by comparing patients with the same posttreatment blood pressure, we mask the effect of one of the two pathways through which the drug operates to bring about recovery.)

When two causal models generate the same statistical data (Figures 6.2(a) and (b) are observationally equivalent), and in one we decide to use the drug yet in the other not to use it, it is obvious that our decision is driven by causal and not by statistical considerations. Some readers might suspect that temporal information is involved in the decision, noting that gender is established before the treatment and blood pressure afterwards. But this is not the case; Figure 6.2(c) shows that F may occur before *or* after C and still the correct decision should remain to consult the combined table (i.e., not to condition on F, as can be seen from the back-door criterion).

We have just demonstrated by example what we already knew in Section 6.1.1 – namely, that every question related to the effect of actions must be decided by causal considerations; statistical information alone is insufficient. Moreover, the question of choosing the correct table on which to base our decision is a special case of the covariate selection problem that was given a general solution in Section 3.3 using causal calculus. Lindley and Novick, on the other hand, stopped short of this realization and attributed the difference between the two examples to a meta-statistical[5] concept called *exchangeability*, first proposed by DeFinetti (1974).

Exchangeability concerns the question of choosing an appropriate reference class, or subpopulation, for making predictions about an individual unit. Insurance companies, for example, would like to estimate the life expectancy of a new customer using mortality records of a class of persons most closely resembling the characteristics of the

[4] The example used in Lindley and Novick (1981) was taken from agriculture, and the causal relationship between C and F was not mentioned, but the structure was the same as in Figure 6.2(b).

[5] By "meta-statistical" I mean a criterion – not itself discernible from statistical data – for judging the adequacy of a certain statistical method.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

new customer. De Finetti gave this question a formal twist by translating judgment about resemblance into judgment of probabilities. According to this criterion, an $(n + 1)$th unit is *exchangeable* in property X, relative to a group of n other units, if the joint probability distribution $P(X_1, \ldots, X_n, X_{n+1})$ is invariant under permutation. To De Finetti, the question of how such invariance can be established was a psychological question of secondary importance; the main point was to cast the target of this psychological exercise in the form of mathematical expression so that it could be communicated and discussed in scientific terms. It is this concept that Lindley and Novick tried to introduce into Simpson's reversal phenomenon and with which they hoped to show that the appropriate subpopulations in the $F =$ gender example are the male and female whereas, in the $F =$ blood pressure example, the whole population of patients should be considered.

Readers of Lindley and Novick's article would quickly realize that, although these authors decorate their discussion with talks of *exchangeability* and *subpopulations*, what they actually do is present informal cause–effect arguments for their intuitive conclusions. Meek and Glymour (1994) keenly observed that the only comprehensible part of Lindley and Novick's discussion of exchangeability is the one based on causal considerations, which suggests that "an explicit account of the interaction of causal beliefs and probabilities is necessary to understand when exchangeability should and should not be assumed" (Meek and Glymour 1994, p. 1013).

This is indeed the case; exchangeability in experimental studies depends on causal understanding of the mechanisms that generate the data. The determination of whether the response of a new unit should be judged by previous response of a group of units is predicated upon the question of whether the experimental conditions to which we contemplate subjecting the new unit are equal to those prevailing when the group was observed. The reason we cannot use the combined table (Figure 6.1(a)) for determining the response of a new patient (of unknown gender) is that the experimental conditions have changed; whereas the group was studied with patients selecting treatment by choice, the new patient will be given treatment by decree, perhaps against his or her natural inclination. A mechanism will therefore be altered in the new experiment, and no judgment of exchangeability is feasible without first making causal assumptions regarding whether the probabilities involved would or would not remain invariant to such alteration. The reason we could use the combined table in the blood pressure example of Figure 6.2(b) is that the altered treatment selection mechanism in that setup is assumed to have no effect on the conditional probability $P(E \mid C)$; that is, C is assumed to be exogenous. (This can clearly be seen in the absence of any back-door path in the graph.)

Note that the same consideration holds if the next patient is a member of the group under study (assuming hypothetically that treatment and effect can be replicated and that the next patient is of unknown gender and identity); a randomly selected sample from a population is not "exchangeable" with that population if we subject the sample to new experimental conditions. Alteration of causal mechanisms must be considered in order to determine whether exchangability holds under the new circumstances. And once causal mechanisms are considered, separate judgment of exchangeability is not needed.

But why did Lindley and Novick choose to speak so elliptically (via exchangeability) when they could have articulated their ideas directly by talking openly about causal

relations? They partially answered this question as follows: "[causality], although widely used, does not seem to be well-defined." One may naturally wonder how exchangeability can be more "well-defined" than the very considerations by which it is judged! The answer can be understood only when we consider the mathematical tools available to statisticians in 1981. When Lindley and Novick wrote that causality is not well defined, what they really meant is that causality cannot be written down in any mathematical form to which they were accustomed. The potentials of path diagrams, structural equations, and Neyman–Rubin notation as mathematical languages were generally unrecognized in 1981, for reasons described in Sections 5.1 and 7.4.3. Indeed, had Lindley and Novick wished to convey their ideas in causal terms, they would have been unable to express mathematically even the simple yet crucial fact that gender is not affected by the drug and a fortiori to derive less obvious truths from that fact.[6] The only formal language with which they were familiar was probability calculus, but as we have seen on several occasions already, this calculus cannot adequately handle causal relationships without the proper extensions.

Fortunately, the mathematical tools that have been developed in the past ten years permit a more systematic and friendly resolution of Simpson's paradox.

6.1.4 A Paradox Resolved (Or: What Kind of Machine Is Man?)

Paradoxes, like optical illusions, are often used by psychologists to reveal the inner workings of the mind, for paradoxes stem from (and amplify) dormant clashes among implicit sets of assumptions. In the case of Simpson's paradox, we have a clash between (i) the assumption that causal relationships are governed by the laws of probability calculus and (ii) the set of implicit assumptions that drive our causal intuitions. The first assumption tells us that the three inequalities in (6.1)–(6.3) are consistent, and it even presents us with a probability model to substantiate the claim (Figure 6.1). The second tells us that no miracle drug can ever exist that is harmful to both males and females and is simultaneously beneficial to the population at large.

To resolve the paradox we must either (a) show that our causal intuition is misleading or incoherent or (b) deny the premise that causal relationships are governed by the laws of standard probability calculus. As the reader surely suspects by now, we will choose the second option; our stance here, as well as in the rest of the book, is that causality is governed by its own logic and that this logic requires a major extension of probability calculus. It still behooves us to explicate the logic that governs our causal intuition and to show, formally, that this logic precludes the existence of such a miracle drug.

The logic of the *do* $(\cdot)$ operator is perfectly suitable for this purpose. Let us first translate the statement that our miracle drug C has a harmful effect on both males and females into formal statements in causal calculus:

[6] Lindley and Novick (1981, p. 50) did try to express this fact in probabilistic notation. But not having the *do* $(\cdot)$ operator at their disposal, they improperly wrote $P(F \mid C)$ instead of $P(F \mid do\ (C))$ and argued unconvincingly that we should equate $P(F \mid C)$ and $P(F)$: "Instead you might judge that the decision to use the treatment or the control is not affected by the unknown sex, so that F and C are independent." Oddly, this decision is also not affected by the unknown blood pressure, and yet, if we write $P(F \mid C) = P(F)$ in the example of Figure 6.2(b), we obtain the wrong result.

$$P(E \mid do(C), F) < P(E \mid do(\neg C), F), \tag{6.4}$$

$$P(E \mid do(C), \neg F) < P(E \mid do(\neg C), \neg F). \tag{6.5}$$

We need to demonstrate that C must be harmful to the population at large; that is, the inequality

$$P(E \mid do(C)) > P(E \mid do(\neg C)) \tag{6.6}$$

must be shown to be inconsistent with what we know about drugs and gender.

Theorem 6.1.1 (Sure-Thing Principle)[7]
An action C that increases the probability of an event E in each subpopulation must also increase the probability of E in the population as a whole, provided that the action does not change the distribution of the subpopulations.

Proof
We will prove Theorem 6.1.1 in the context of our example, where the population is partitioned into males and females; generalization to multiple partitions is straightforward. In this context, we need to prove that the reversal in the inequalities of (6.4)–(6.6) is inconsistent with the assumption that drugs have no effect on gender:

$$P(F \mid do(C)) = P(F \mid do(\neg C)) = P(F). \tag{6.7}$$

Expanding $P(E \mid do(C))$ and using (6.7) yields

$$\begin{aligned} P(E \mid do(C)) &= P(E \mid do(C), F)\, P(F \mid do(C)) \\ &\quad + P(E \mid do(C), \neg F)\, P(\neg F \mid do(C)) \\ &= P(E \mid do(C), F)\, P(F) + P(E \mid do(C), \neg F)\, P(\neg F). \end{aligned} \tag{6.8}$$

Similarly, for $do(\neg C)$ we obtain

$$\begin{aligned} P(E \mid do(\neg C)) &= P(E \mid do(\neg C), F)\, P(F) \\ &\quad + P(E \mid do(\neg C), \neg F)\, P(\neg F). \end{aligned} \tag{6.9}$$

Since every term on the right-hand side of (6.8) is smaller than the corresponding term in (6.9), we conclude that

[7] Savage (1954, p. 21) proposed the sure-thing principle as a basic postulate of preferences (on actions), tacitly assuming the no-change provision in the theorem. Blyth (1972) used this omission to devise an apparent counterexample. Theorem 6.1.1 shows that the sure-thing principle need not be stated as a separate postulate – it follows logically from the semantics of actions as modifiers of structural equations (or mechanisms). See Gibbard and Harper (1976) for a counterfactual analysis. Note that the no-change provision is probabilistic; it permits the action to change the classification of individual units as long as the relative sizes of the subpopulations remain unaltered.

$$P(E \mid do\,(C)) < P\,(E \mid do\,(\neg C)),$$

proving Theorem 6.1.1. □

We thus see where our causal intuition comes from: an obvious but crucial assumption in our intuitive logic has been that drugs do not influence gender. This explains why our intuition changes so drastically when F is interpreted as an intermediate event affected by the drug, as in Figure 6.2(b). In this case, our intuitive logic tells us that it is perfectly consistent to find a drug satisfying the three inequalities of (6.4)–(6.6) and, moreover, that it would be inappropriate to adjust for F. If F is affected by the C, then (6.8) cannot be derived, and the difference $P(E \mid do\,(C)) - P\,(E \mid do\,(\neg C))$ may be positive or negative, depending on the relative magnitudes of $P\,(F \mid do\,(C))$ and $P(F \mid do\,(\neg C))$. Provided C and E have no common cause, we should then assess the efficacy of C directly from the combined table (equation (6.1)) and not from the F-specific tables (equations (6.2)–(6.3)).

Note that nowhere in our analysis have we assumed either that the data originate from a randomized study (i.e., $P(E \mid do(C)) = P(E \mid C)$) or from a balanced study (i.e., $P\,(C \mid F) = P\,(C \mid \neg F)$). On the contrary, given the tables of Figure 6.1, our causal logic accepts gracefully that we are dealing with an unbalanced study but nevertheless refuses to accept the consistency of (6.4)–(6.6). People, likewise, can see clearly from the tables that the males were more likely to take the drug than the females; still, when presented with the reversal phenomenon, people are "shocked" to discover that differences in recovery rates can be reversed by combining tables.

The conclusions we may draw from these observations are that humans are generally oblivious to rates and proportions (which are transitory) and that they constantly search for causal relations (which are invariant). Once people interpret proportions as causal relations, they continue to process those relations by causal calculus and not by the calculus of proportions. Were our minds governed by the calculus of proportions, Figure 6.1 would have evoked no surprise at all, and Simpson's paradox would never have generated the attention that it did.

6.2 WHY THERE IS NO STATISTICAL TEST FOR CONFOUNDING, WHY MANY THINK THERE IS, AND WHY THEY ARE ALMOST RIGHT

6.2.1 Introduction

Confounding is a simple concept. If we undertake to estimate the effect[8] of one variable (X) on another (Y) by examining the statistical association between the two, we ought to ensure that the association is not produced by factors other than the effect under study. The presence of spurious association – due, for example, to the influence of extraneous variables – is called *confounding* because it tends to confound our reading and to

[8] We will confine the use of the terms "effect," "influence," and "affect" to their causal interpretations; the term "association" will be set aside for statistical dependencies.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

bias our estimate of the effect studied. Conceptually, therefore, we can say that X and Y are confounded when there is a third variable Z that influences both X and Y; such a variable is then called a *confounder* of X and Y.

As simple as this concept is, it has resisted formal treatment for decades, and for good reason: The very notions of "effect" and "influence" – relative to which "spurious association" must be defined – have resisted mathematical formulation. The empirical definition of effect as an association that *would* prevail in a controlled randomized experiment cannot easily be expressed in the standard language of probability theory, because that theory deals with static conditions and does not permit us to predict, even from a full specification of a population density function, what relationships would prevail if conditions were to change – say, from observational to controlled studies. Such predictions require extra information in the form of causal or counterfactual assumptions, which are not discernible from density functions (see Sections 1.3 and 1.4). The $do(\cdot)$ operator used in this book was devised specifically for distinguishing and managing this extra information.

These difficulties notwithstanding, epidemiologists, biostatisticians, social scientists, and economists[9] have made numerous attempts to define confounding in statistical terms, partly because statistical definitions – free of theoretical terms of "effect" or "influence" – can be expressed in conventional mathematical form and partly because such definitions may lead to practical tests of confounding and thereby alert investigators to possible bias and need for adjustment. These attempts have converged in the following basic criterion.

Associational Criterion

Two variables X and Y are not confounded if and only if every variable Z that is not affected by X is either

(U_1) *unassociated with X or*

(U_2) *unassociated with Y, conditional on X.*

This criterion, with some variations and derivatives (often avoiding the "only if" part), can be found in almost every epidemiology textbook (Schlesselman 1982; Rothman 1986; Rothman and Greenland 1998) and in almost every article dealing with confounding. In fact, the criterion has become so deeply entrenched in the literature that authors (e.g., Gail 1986; Hauck et al. 1991; Becher 1992; Steyer et al. 1996) often take it to be the *definition* of no-confounding, forgetting that ultimately confounding is useful only so far as it tells us about effect bias.[10]

The purpose of this and the next section is to highlight several basic limitations of the associational criterion and its derivatives. We will show that the associational criterion

[9] In econometrics, the difficulties have focused on the notion of "exogeneity" (Engle et al. 1983; Leamer 1985; Aldrich 1993), which stands essentially for "no confounding" (see Section 5.4.3).

[10] Hauck et al. (1991) dismiss the effect-based definition of confounding as "philosophic" and consider a difference between two measures of association to be a "bias." Grayson (1987) even goes so far as to state that the change-in-parameter method, a derivative of the associational criterion, is the only fundamental definition of confounding (see Greenland et al. 1989 for critiques of Grayson's position).

neither ensures unbiased effect estimates nor follows from the requirement of unbiasedness. After demonstrating, by examples, the absence of logical connections between the statistical and the causal notions of confounding, we will define a stronger notion of unbiasedness, called "stable" unbiasedness, relative to which a modified statistical criterion will be shown necessary and sufficient. The necessary part will then yield a practical test for stable unbiasedness that, remarkably, does not require knowledge of all potential confounders in a problem. Finally, we will argue that the prevailing practice of substituting statistical criteria for the effect-based definition of confounding is not entirely misguided, because stable unbiasedness is in fact (i) what investigators have been (and perhaps should be) aiming to achieve and (ii) what statistical criteria can test.

6.2.2 Causal and Associational Definitions

In order to facilitate the discussion, we shall first cast the causal and statistical definitions of no-confounding in mathematical forms.[11]

Definition 6.2.1 (No-Confounding; Causal Definition)
Let M be a causal model of the data-generating process – that is, a formal description of how the value of each observed variable is determined. Denote by $P(y \mid do(x))$ *the probability of the response event* $Y = y$ *under the hypothetical intervention* $X = x$*, calculated according to M. We say that X and Y are not confounded in M if and only if*

$$P(y \mid do(x)) = P(y \mid x), \text{ or } P(x \mid do(y)) = P(x \mid y) \tag{6.10}$$

for all x and y in their respective domains, where $P(y \mid x)$ *is the conditional probability generated by M. If* (6.10) *holds, we say that* $P(y \mid x)$ *is unbiased.*

For the purpose of our discussion here, we take this causal definition as the meaning of the expression "no confounding." The probability $P(y \mid do(x))$ was defined in Chapter 3 (Definition 3.2.1, also abbreviated $P(y \mid \hat{x})$); it may be interpreted as the conditional probability $P^*(Y = y \mid X = x)$ corresponding to a controlled experiment in which X is randomized. We recall that this probability can be calculated from a causal model M either directly, by simulating the intervention $do(X = x)$, or (if $P(x, s) > 0$) via the adjustment formula (equation (3.19))

$$P(y \mid do(x)) = \sum_s P(y \mid x, s)\, P(s),$$

where S stands for any set of variables, observed as well as unobserved, that satisfy the back-door criterion (Definition 3.3.1). Equivalently, $P(y \mid do(x))$ can be written $P(Y(x) = y)$, where $Y(x)$ is the potential-outcome variable as defined in (3.51) or in

[11] For simplicity, we will limit our discussion to unadjusted confounding; extensions involving measurement of auxiliary variables are straightforward and can be obtained from Section 3.3. We also use the abbreviated expression "X and Y are not confounded," though "the effect of X on Y is not confounded" is more exact.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

Rubin (1974). We bear in mind that the operator *do* (·), and hence also effect estimates and confounding, must be defined relative to a specific causal or data-generating model *M* because these notions are not statistical in character and cannot be defined in terms of joint distributions.

Definition 6.2.2 (No-Confounding; Associational Criterion)
Let T be the set of variables in a problem that are not affected by X. We say that X and Y are not confounded in the presence of T if each member Z of T satisfies at least one of the following conditions:

(U_1) *Z is not associated with X* (*i.e.*, $P(x \mid z) = P(x)$);
(U_2) *Z is not associated with Y, conditional on X* (*i.e.*, $P(y \mid z, x) = P(y \mid x)$).

Conversely, X and Y are said to be confounded if any member Z of T violates both (U_1) *and* (U_2).

Note that the associational criterion in Definition 6.2.2 is not purely statistical in that it invokes the predicate "affected by," which is not discernible from probabilities but rests instead on causal information. This exclusion of variables that are affected by treatments (or exposures) is unavoidable and has long been recognized as a necessary judgmental input to every analysis of treatment effect in observational and experimental studies alike (Cox 1958, p. 48; Greenland and Neutra 1980). We shall assume throughout that investigators possess the knowledge required for distinguishing variables that are affected by the treatment *X* from those that are not. We shall then explore what additional causal knowledge is needed, if any, for establishing a test of confounding.

6.3 HOW THE ASSOCIATIONAL CRITERION FAILS

We will say that a criterion for no-confounding is *sufficient* if it never errs when it classifies a case as no-confounding and *necessary* if it never errs when it classifies a case as confounding. There are several ways that the associational criterion of Definition 6.2.2 fails to match the causal criterion of Definition 6.2.1. Failures with respect to sufficiency and necessity will be addressed in turn.

6.3.1 Failing Sufficiency via Marginality

The criterion in Definition 6.2.2 is based on testing each element of *T* individually. A situation may well be present where two factors, Z_1 and Z_2, jointly confound *X* and *Y* (in the sense of Definition 6.2.2) and yet each factor separately satisfies (U_1) or (U_2). This may occur because statistical independence between *X* and individual members of *T* does not guarantee the independence of *X* and groups of variables taken from *T*. For example, let Z_1 and Z_2 be the outcomes of two independent fair coins, each affecting both *X* and *Y*. Assume that *X* occurs when Z_1 and Z_2 are equal and that *Y* occurs whenever Z_1 and Z_2 are unequal. Clearly, *X* and *Y* are highly confounded by the pair $T = (Z_1, Z_2)$; they are, in fact, perfectly correlated (negatively) without causally affecting

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

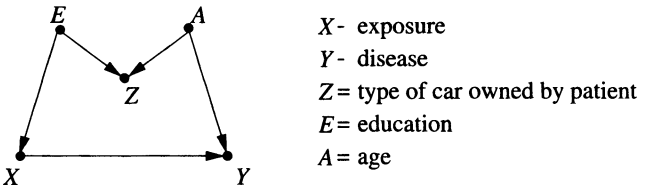

Figure 6.3 X and Y are not confounded, though Z is associated with both.

each other. Yet neither Z_1 nor Z_2 is associated with either X or Y; discovering the outcome of any one coin does not change the probability of X (or of Y) from its initial value of $\frac{1}{2}$.

An attempt to remedy Definition 6.2.2 by replacing Z with arbitrary subsets of T in (U_1) and (U_2) would be much too restrictive, because the set of *all* causes of X and Y, when treated as a group, would almost surely fail the tests of (U_1) and (U_2). In Section 6.5.2 we identify the subsets that should replace Z in (U_1) and (U_2) if sufficiency is to be restored.

6.3.2 Failing Sufficiency via Closed-World Assumptions

By "closed-world" assumption I mean the assumption that our model accounts for all relevant variables and, specifically to Definition 6.2.2, that the set T of variables consists of *all* potential confounders in a problem. In order to correctly classify every case of no-confounding, the associational criterion requires that condition (U_1) or (U_2) be satisfied for every potential confounder Z in a problem. In practice, since investigators can never be sure whether a given set T of potential confounders is complete, the associational criterion will falsely classify certain confounded cases as unconfounded.

This limitation actually implies that any statistical test whatsoever is destined to be insufficient. Since practical tests always involve proper subsets of T, the most we can hope to achieve by statistical means is *necessity* – that is, a test that would correctly label cases as confounding when criteria such as (U_1) and (U_2) are violated by an arbitrary subset of T. This prospect too is not fulfilled by Definition 6.2.2, as we now demonstrate.

6.3.3 Failing Necessity via Barren Proxies

Example 6.3.1 Imagine a situation where exposure (X) is influenced by a person's education (E), disease (Y) is influenced by both exposure and age (A), and car type (Z) is influenced by both age (A) and education (E). These relationships are shown schematically in Figure 6.3.

The car-type variable (Z) violates the two conditions in Definition 6.2.2 because: (1) car type is indicative of education and hence is associated with the exposure variable; and (2) car type is indicative of age and hence is associated with the disease among the exposed and the nonexposed. However, in this example the effect of X on Y is not confounded; the type of car owned by a person has no effect on either exposure or disease and is merely one among many irrelevant properties that are associated with both via intermediaries. The analysis of Chapter 3 establishes that,

indeed, (6.10) is satisfied in this model[12] and that, moreover, adjustment for Z would generally yield a biased result:

$$\sum_z P(Y = y \mid X = x, Z = z)\, P(Z = z) \neq P(Y = y \mid do(x)).$$

Thus we see that the traditional criterion based on statistical association fails to identify an unconfounded effect and would tempt one to adjust for the wrong variable. This failure occurs whenever we apply (U_1) and (U_2) to a variable Z that is a *barren proxy* – that is, a variable that has no influence on X or Y but is a proxy for factors that do have such influence.

Readers may not consider this failure to be too serious, because experienced epidemiologists would rarely regard a variable as a confounder unless it is suspect of having some influence on either X or Y. Nevertheless, adjustment for proxies is a prevailing practice in epidemiology and should be done with great caution (Greenland and Neutra 1980; Weinberg 1993). To regiment this caution, the associational criterion must be modified to exclude barren proxies from the test set T. This yields the following modified criterion in which T consists only of variables that (causally) influence Y (possibly through X).

Definition 6.3.2 (No-Confounding; Modified Associational Criterion)
Let T be the set of variables in a problem that are not affected by X but may potentially affect Y. We say that X and Y are unconfounded by the presence of T if and only if every member Z of T satisfies either (U_1) *or* (U_2) *of Definition* 6.2.2.

Stone (1993) and Robins (1997) proposed alternative modifications of Definition 6.2.2 that avoid the problems created by barren proxies without requiring one to judge whether a variable has an effect on Y. Instead of restricting the set T to potential causes of Y, we let T remain the set of *all* variables unaffected by X,[13] requiring instead that T be composed of two disjoint subsets, T_1 and T_2, such that

(U_1^*) T_1 is unassociated with X *and*
(U_2^*) T_2 is unassociated with Y given X and T_1.

In the model of Figure 6.3, for instance, conditions $((U_1^*)$ and (U_2^*) are satisfied by the choice $T_1 = A$ and $T_2 = \{Z, E\}$, because (using the d-separation test) A is independent of X and $\{E, Z\}$ is independent of Y, given $\{X, A\}$.

This modification of the associational criterion further rectifies the problem associated with marginality (see Section 6.3.1) because (U_1^*) and (U_2^*) treat T_1 and T_2 as compound

[12] Because the (back-door) path $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$ is blocked by the colliding arrows at Z (see Definition 3.3.1).

[13] Alternatively, T can be confined to any set S of variables sufficient for control of confounding:

$$P(y \mid do(x)) = \sum_s P(y \mid x, s) P(s).$$

Again, however, we can never be sure if the measured variables in the model contain such a set, or which of T's subsets possess this property.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

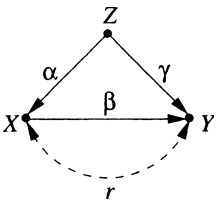

Figure 6.4 Z is associated with both X and Y, yet the effect of X on Y is not confounded (when $r = -\alpha\gamma$).

variables. However, the modification falls short of restoring necessity. Because the set $T = (T_1, T_2)$ must include *all* variables unaffected by X (see note 13) and because practical tests are limited to proper subsets of T, we cannot conclude that confounding is present solely upon the failure of (U_1^*) and (U_2^*), as specified in Section 6.3.2. This criterion too is thus inadequate as a basis for practical detection of confounding.

We now discuss another fundamental limitation on our ability to detect confounding by statistical means.

6.3.4 Failing Necessity via Incidental Cancellations

Here we present a case that is devoid of barren proxies and in which the effect of X on Y (i) is not confounded in the sense of (6.10) but (ii) is confounded according to the modified associational criterion of Definition 6.3.2.

Example 6.3.3 Consider a causal model defined by the linear equations

$$x = \alpha z + \varepsilon_1, \tag{6.11}$$

$$y = \beta x + \gamma z + \varepsilon_2, \tag{6.12}$$

where ε_1 and ε_2 are correlated unmeasured variables with $\text{cov}(\varepsilon_1, \varepsilon_2) = r$ and where Z is an exogenous variable that is uncorrelated with ε_1 or ε_2. The diagram associated with this model is depicted in Figure 6.4. The effect of X on Y is quantified by the path coefficient β, which gives the rate of change of $E(Y \mid do(x))$ per unit change in x.[14]

It is not hard to show (assuming standardized variables) that the regression of Y on X gives

$$y = (\beta + r + \alpha\gamma)\, x + \varepsilon,$$

where $\text{cov}(x, \varepsilon) = 0$. Thus, whenever the equality $r = -\alpha\gamma$ holds, the regression coefficient of $r_{YX} = \beta + r + \alpha\gamma$ is an unbiased estimate of β, meaning that the effect of X on Y is unconfounded (no adjustment is necessary). Yet the associational conditions (U_1) and (U_2) are both violated by the variable Z; Z is associated with X (if $\alpha \neq 0$) and conditionally associated with Y, given X (except for special values of γ for which $\rho_{yz \cdot x} = 0$).

[14] See Sections 3.5–3.6 or (5.24) in Section 5.4.1.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

This example demonstrates that the condition of unbiasedness (Definition 6.2.1) does not imply the modified criterion of Definition 6.3.2. The associational criterion might falsely classify some unconfounded situations as confounded and, worse yet, adjusting for the false confounder (Z in our example) will introduce bias into the effect estimate.[15]

6.4 STABLE VERSUS INCIDENTAL UNBIASEDNESS

6.4.1 Motivation

The failure of the associational criterion in the previous example calls for a reexamination of the notion of confounding and unbiasedness as defined in (6.10). The reason that X and Y were classified as unconfounded in Example 6.3.3 was that, by setting $r = -\alpha\gamma$, we were able to make the spurious association represented by r *cancel* the one mediated by Z. In practice, such perfect cancellation would be an incidental event specific to a peculiar combination of study conditions, and it would not persist when the parameters of the problem (i.e., α, γ, and r) undergo slight changes – say, when the study is repeated in a different location or at a different time. In contrast, the condition of no-confounding found in Example 6.3.1 does not exhibit such volatility. In this example, the unbiasedness expressed in (6.10) would continue to hold regardless of the strength of connection between education and exposure and regardless on how education and age influence the type of car that a patient owns. We call this type of unbiasedness *stable*, since it is robust to change in parameters and remains intact as long as the configuration of causal connections in the model remains the same.

In light of this distinction between stable and incidental unbiasedness, we need to reexamine whether we should regard a criterion as inadequate if it misclassifies (as confounded) cases that are rendered unconfounded by mere incidental cancellation and, more fundamentally, whether we should insist on including such peculiar cases in the definition of unbiasedness (given the precarious conditions under which (6.10) would be satisfied in these cases). Although answers to these questions are partly a matter of choice, there is ample evidence that our intuition regarding confounding is driven by considerations of stable unbiasedness, not merely incidental ones. How else can we explain why generations of epidemiologists and biostatisticians would advocate confounding criteria that fail in cases involving incidental cancellation? On the pragmatic side, failing to detect situations of incidental unbiasedness should not introduce appreciable error in observational studies because those situations are short-lived and are likely to be refuted by subsequent studies, under slightly different conditions.[16]

Assuming that we are prepared to classify as unbiased only cases in which unbiasedness remains robust to changes in parameters, two questions remain: (1) How can we give this new notion of "stable unbiasedness" a formal, nonparametric formulation? (2) Are practical statistical criteria available for testing stable unbiasedness? Both questions can be answered using structural models.

---

[15] Note that the Stone–Robins modifications of Definition 6.3.2 would also fail in this example, unless we can measure the factors responsible for the correlation between ε_1 and ε_2.

[16] As we have seen in Example 6.3.3, any statistical test capable of recognizing such cases would require measurement of *all* variables in T.

Chapter 3 describes a graphical criterion, called the "back-door criterion," for identifying conditions of unbiasedness in a causal diagram.[17] In the simple case of no adjustment (for measured covariates), the criterion states that X and Y are unconfounded if every path between X and Y that contains an arrow pointing into X must also contain a pair of arrows pointing head-to-head (as in Figure 6.3); this criterion is valid whenever the missing links in the diagram represent absence of causal connections among the corresponding variables. Because the causal assumptions embedded in the missing links are so explicit, the back-door criterion has two remarkable features. First, no statistical information is needed; the topology of the diagram suffices for reliably determining whether an effect is unconfounded (in the sense of Definition 6.2.1) and whether an adjustment for a set of variables is sufficient for removing confounding when one exists. Second, any model that meets the back-door criterion would in fact satisfy (6.10) for an infinite class of models (or situations), each generated by assigning different parameters to the causal connections in the diagram.

To illustrate, consider the diagram depicted in Figure 6.3. The back-door criterion will identify the pair (X, Y) as unconfounded, because the only path ending with an arrow into X is the one traversing (X, E, Z, A, Y), and this path contains two arrows pointing head-to-head at Z. Moreover, since the criterion is based only on graphical relationships, it is clear that (X, Y) will continue to be classified as unconfounded regardless of the strength or type of causal relationships that are represented by the arrows in the diagram. In contrast, consider Figure 6.4 in Example 6.3.3, where two paths end with arrows into X. Since none of these paths contains head-to-head arrows, the back-door criterion will fail to classify the effect of X on Y as unconfounded, acknowledging that an equality $r = -\alpha\gamma$ (if it prevails) would not represent a stable case of unbiasedness.

The vulnerability of the back-door criterion to causal assumptions can be demonstrated in the context of Figure 6.3. Assume the investigator suspects that variable Z (car type) has some influence on the outcome variable Y. This would amount to adding an arrow from Z to Y in the diagram, classifying the situation as confounded, and suggesting an adjustment for E (or $\{A, Z\}$). Yet no adjustment is necessary if, owing to the specific experimental conditions in the study, Z has in fact no influence on Y. It is true that the adjustment suggested by the back-door criterion would introduce no bias, but such adjustment could be costly if it calls for superfluous measurements in a no-confounding situation.[18] The added cost is justified in light of (i) the causal information at hand (i.e., that Z may potentially influence Y) and (ii) our insistence on ensuring stable unbiasedness – that is, avoiding bias in all situations compatible with the information at hand.

[17] A gentle introduction to applications of the back-door criterion in epidemiology can be found in Greenland et al. (1999a).

[18] On the surface, it appears as though the Stone–Robins criterion would correctly recognize the absence of confounding in this situation, since it is based on associations that prevail in the probability distribution that actually generates the data (according to which $\{E, Z\}$ should be independent of Y, given $\{A, X\}$). However, these associations are of no help in deciding whether certain measurements can be *avoided*; such decisions must be made prior to gathering the data and must rely therefore on subjective assumptions about the disappearance of conditional associations. Such assumptions are normally supported by causal, not associational, knowledge (see Section 1.3).

6.4.2 Formal Definitions

To formally distinguish between *stable* and *incidental* unbiasedness, we use the following general definition.

Definition 6.4.1 (Stable Unbiasedness)
Let A be a set of assumptions (*or restrictions*) *on the data-generating process, and let* C_A *be a class of causal models satisfying A. The effect estimate of X on Y is said to be* stably unbiased *given A if* $P(y \mid do(x)) = P(y \mid x)$ *holds in every model M in* C_A. *Correspondingly, we say that the pair* (*X, Y*) *is* stably unconfounded *given A.*

The assumptions commonly used to specify causal models can be either parametric or topological. For example, the structural equation models used in the social sciences and economics are usually restricted by the assumptions of linearity and normality. In this case, C_A would consist of all models created by assigning different values to the unspecified parameters in the equations and in the covariance matrix of the error terms. Weaker, nonparametric assumptions emerge when we specify merely the topological structure of the causal diagram but let the error distributions and the functional form of the equations remain undetermined. We now explore the statistical ramifications of these nonparametric assumptions.

Definition 6.4.2 (Structurally Stable No-Confounding)
Let A_D *be the set of assumptions embedded in a causal diagram D. We say that X and Y are stably unconfounded given* A_D *if* $P(y \mid do(x)) = P(y \mid x)$ *holds in every parameterization of D. By "parameterization" we mean an assignment of functions to the links of the diagram and prior probabilities to the background variables in the diagram.*

Explicit interpretation of the assumptions embedded in a causal diagram are given in Chapters 3 and 5. Put succinctly, if *D* is the diagram associated with the causal model, then:

1. every missing arrow (between, say, *X* and *Y*) represents the assumption that *X* has no effect on *Y* once we intervene and hold the parents of *Y* fixed;
2. every missing bidirected link between *X* and *Y* represents the assumption that there are no common causes for *X* and *Y*, except those shown in *D*.

Whenever the diagram *D* is acyclic, the back-door criterion provides a necessary and sufficient test for stable no-confounding, given A_D. In the simple case of no adjustment for covariates, the criterion reduces to the nonexistence of a common ancestor, observed or latent, of *X* and *Y*.[19] Thus, we have our next theorem.

[19] The colloquial term "common ancestors" should exclude nodes that have no other connection to *Y* except through *X* (e.g., node *E* in Figure 6.3) and include latent nodes for correlated errors. In the diagram of Figure 6.4, for example, *X* and *Y* are understood to have two common ancestors; the first is *Z* and the second is the (implicit) latent variable responsible for the double-arrowed arc between *X* and *Y* (i.e., the correlation between ε_1 and ε_2).

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

Theorem 6.4.3 (Common-Cause Principle)
Let A_D be the set of assumptions embedded in an acyclic causal diagram D. Variables X and Y are stably unconfounded given A_D if and only if X and Y have no common ancestor in D.

Proof
The "if" part follows from the validity of the back-door criterion (Theorem 3.3.2). The "only if" part requires the construction of a specific model in which (6.10) is violated whenever X and Y have a common ancestor in D. This is easily done using linear models and Wright's rules for path coefficients. □

Theorem 6.4.3 provides a necessary and sufficient condition for stable no-confounding without invoking statistical data, since it relies entirely on the information embedded in the diagram. Of course, the diagram itself has statistical implications that can be tested (Sections 1.2.3 and 5.2.1), but those tests do not specify the diagram uniquely (see Chapter 2 and Section 5.2.3).

Suppose, however, that we do not possess all the information required for constructing a causal diagram and instead know merely for each variable Z whether it is safe to assume that Z has no effect on Y and whether X has no effect on Z. The question now is whether this more modest information, together with statistical data, is sufficient to qualify or disqualify a pair (X, Y) as stably unconfounded. The answer is positive.

6.4.3 Operational Test for Stable No-Confounding

Theorem 6.4.4 (Criterion for Stable No-Confounding)
Let A_Z denote the assumptions that (i) *the data are generated by some* (*unspecified*) *acyclic model M and* (ii) *Z is a variable in M that is unaffected by X but may possibly affect Y.*[20] *If both of the associational criteria* (U_1) *and* (U_2) *of Definition* 6.2.2 *are violated, then* (X, Y) *are not stably unconfounded given* A_Z.

Proof
Whenever X and Y are stably unconfounded, Theorem 6.4.3 rules out the existence of a common ancestor of X and Y in the diagram associated with the underlying model. The absence of a common ancestor, in turn, implies the satisfaction of either (U_1) or (U_2) whenever Z satisfies A_Z. This is a consequence of the d-separation rule (Section 1.2.3) for reading the conditional independence relationships entailed by a diagram.[21] □

Theorem 6.4.4 implies that the traditional associational criteria (U_1) and (U_2) could be used in a simple operational test for stable no-confounding, a test that does not require us to know the causal structure of the variables in the domain or even to enumerate the set of relevant variables. Finding just *any* variable Z that satisfies A_Z and violates (U_1)

[20] By "possibly affecting Y" we mean: A_Z does not contain the assumption that Z does not affect Y. In other words, the diagram associated with M must contain a directed path from Z to Y.
[21] It also follows from Theorem 7(a) in Robins (1997).

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

and (U_2) permits us to disqualify (X, Y) as stably unconfounded (though (X, Y) may be incidentally unconfounded in the particular experimental conditions prevailing in the study).

Theorem 6.4.4 communicates a formal connection between statistical associations and confounding that is not based on the closed-world assumption.[22] It is remarkable that the connection can be formed under such a weak set of added assumptions: the qualitative assumption that a variable may have influence on Y and is not affected by X suffices to produce a necessary statistical test for stable no-confounding.

6.5 CONFOUNDING, COLLAPSIBILITY, AND EXCHANGEABILITY

6.5.1 Confounding and Collapsibility

Theorem 6.4.4 also establishes a formal connection between confounding and "collapsibility" – a criterion under which a measure of association remains invariant to the omission of certain variables.

Definition 6.5.1 (Collapsibility)
Let g $[P(x, y)]$ *be any functional*[23] *that measures the association between* Y *and* X *in the joint distribution* $P(x, y)$. *We say that* g *is* collapsible *on a variable* Z *if*

$$E_z g[P(x, y \mid z)] = g[P(x, y)].$$

It is not hard to show that if g stands for any linear functional of $P(y \mid x)$ – for example, the risk difference $P(y \mid x_1) - P(y \mid x_2)$ – then collapsibility holds whenever Z is either unassociated with X or unassociated with Y given X. Thus, any violation of collapsibility implies violation of the two statistical criteria of Definition 6.2.2, and that is probably why many believed noncollapsibility to be intimately connected with confounding. However, the examples in this chapter demonstrate that violation of these two conditions is neither sufficient nor necessary for confounding. Thus, noncollapsibility and confounding are in general two distinct notions; neither implies the other.

Some authors tend to believe that this distinction is a peculiar property of nonlinear effect measures g, such as the odds or likelihood ratios, and that "when the effect measure is an expectation over population units, confounding and noncollapsibility are algebraically equivalent" (Greenland 1998, p. 906). This chapter shows that confounding and noncollapsibility need not correspond even in linear functionals. For example, the effect measure $P(y \mid x_1) - P(y \mid x_2)$ (the risk difference) is not collapsible over Z in Figure 6.3 (for almost every parameterization of the graph) and yet the effect measure is unconfounded (for every parameterization).

---

[22] I am not aware of another such connection in the literature.

[23] A *functional* is an assignment of a real number to any function from a given set of functions. For example, the mean $E(X) = \Sigma_x x P(x)$ is a functional, since it assigns a real number $E(X)$ to each probability function $P(x)$.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

The logical connection between confounding and collapsibility is formed through the notion of *stable no-confounding*, as formulated in Definition 6.4.2 and Theorem 6.4.4. Because any violation of collapsibility means violation of (U_1) and (U_2) in Definition 6.2.2, it also implies (by Theorem 6.4.4) violation of stable unbiasedness (or stable no-confounding). Thus we can state the following corollary.

Corollary 6.5.2 (Stable No-Confounding Implies Collapsibility)
Let Z be any variable that is not affected by X and that may possibly affect Y. Let g $[P(x, y)]$ *be any linear functional that measures the association between X and Y. If g is not collapsible on Z, then X and Y are not stably unconfounded.*

This corollary provides a rationale for the widespread practice of testing confoundedness by the change-in-parameter method, that is, labeling a variable Z a confounder whenever the "crude" measure of association, $g[P(x, y)]$, is not equal to the Z-specific measures of association averaged over the levels of Z (Breslow and Day 1980; Kleinbaum et al. 1982; Yanagawa 1984; Grayson 1987). Theorem 6.4.4 suggests that the intuitions responsible for this practice were shaped by a quest for a stable condition of no-confounding, not merely an incidental one. Moreover, condition A_Z in Theorem 6.4.4 justifies a requirement made by some authors that a confounder must be a causal determinant of, and not merely associated with, the outcome variable Y.

6.5.2 Confounding versus Confounders

The focus of our discussion in this chapter has been the phenomenon of confounding, which we equated with that of effect bias (Definition 6.2.1). Much of the literature on this topic has been concerned with the presence or absence of *confounders*, presuming that some variables possess the capacity to confound and some do not. This notion may be misleading if interpreted literally, and caution should be exercised before we label a variable as a confounder.

Rothman and Greenland (1998, p. 120), for example, offer this definition: "The extraneous factors responsible for difference in disease frequency between the exposed and unexposed are called *confounders*"; they go on to state that: "In general, a confounder must be associated with both the exposure under study and the disease under study to be confounding" (p. 121). Rothman and Greenland qualify their statement with "In general," and for good reason: We have seen (in the two-coin example of Section 6.3.1) that each individual variable in a problem can be *unassociated* with both the exposure (X) and the disease (Y) under study and still the effect of X on Y remains confounded. A similar situation can also be seen in the linear model depicted in Figure 6.5. Although Z is clearly a confounder for the effect of X on Y and must therefore be controlled, the association between Z and Y may actually vanish (at each level of X); similarly, the association between Z and X may vanish. This can occur if the indirect association mediated by the path $Z \leftarrow A \rightarrow Y$ happens to cancel the direct association carried by the arrow $Z \rightarrow Y$. This cancellation does not imply the absence of confounding, because the path $X \leftarrow E \rightarrow Z \rightarrow Y$ is unblocked while $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$ is blocked. Thus, Z is a confounder that is not associated with the disease (Y).

The intuition behind Rothman and Greenland's statement just quoted can be explicated formally through the notion of stability: a variable that is *stably* unassociated with

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

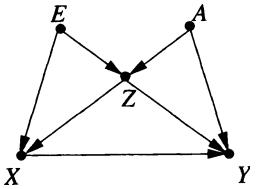

Figure 6.5 Z may be unassociated with both X and Y and still be a confounder (i.e., a member of every sufficient set).

either X or Y can safely be excluded from adjustment. Alternatively, Rothman and Greenland's statement can be supported (without invoking stability) by using the notion of a *nontrivial sufficient set* (Section 3.3) – a set of variables for which adjustment will remove confounding bias. It can be shown (see the end of this section) that each such set S, taken as a unit, must indeed be associated with X and be conditionally associated with Y, given X. Thus, Rothman and Greenland's condition is valid for nontrivial sufficient (i.e., admissible) sets but not for the individual variables in the set.

The practical ramifications of this condition are as follows. If we are given a set S of variables that is claimed to be sufficient (for removing bias by adjustment), then that claim can be given a necessary statistical test: S as a compound variable must be associated both with X and with Y (given X). In Figure 6.5, for example, $S_1 = \{A, Z\}$ and $S_2 = \{E, Z\}$ are sufficient and nontrivial; both must satisfy the condition stated.

Note, however, that although this test can screen out some obviously bad sets S claimed to be sufficient, it has nothing to do with sufficiency or confounding; it merely tests for nontriviality, i.e., that adjusting for S would change the association between X and Y. When we find a nontrivial set S, we still cannot be sure whether the association was unbiased to start with (as in Figure 6.3) or that it turned unbiased after the adjustment.

Proof of Necessity

To prove that (U_1) and (U_2) must be violated whenever Z stands for a nontrivial sufficient set S, consider the case where X has no effect on Y. In this case, confounding amounts to a nonvanishing association between X and Y. A well-known property of conditional independence, called *contraction* (Section 1.1.5), states that (U_1), $X \perp\!\!\!\perp S$, together with sufficiency, $X \perp\!\!\!\perp Y \mid S$, implies violation of nontriviality, $X \perp\!\!\!\perp Y$:

$$X \perp\!\!\!\perp S \;\;\&\;\; X \perp\!\!\!\perp Y \mid S \implies X \perp\!\!\!\perp Y.$$

Likewise, another property of conditional independence, called *intersection,* states that (U_2), $S \perp\!\!\!\perp Y \mid X$, together with sufficiency, $X \perp\!\!\!\perp Y \mid S$, also implies violation of nontriviality, $X \perp\!\!\!\perp Y$.

$$S \perp\!\!\!\perp Y \mid X \;\;\&\;\; X \perp\!\!\!\perp Y \mid S \implies X \perp\!\!\!\perp Y.$$

Thus, both (U_1) and (U_2) must be violated by any nontrivial sufficient set S.

Note, however, that intersection holds only for strictly positive probability distributions, which means that the Rothman–Greenland condition may be violated if deterministic

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

relationships hold among some variables in a problem. This can be seen from a simple example in which both X and Y stand in a one-to-one functional relationship to a third variable, Z. Clearly, Z is a nontrivial sufficient set yet is not associated with Y given X; once we know the value of X, the probability of Y is determined, and would no longer change with learning the value of Z.

6.5.3 Exchangeability versus Structural Analysis of Confounding

Students of epidemiology complain bitterly about the confusing way in which the fundamental concept of confounding has been treated in the literature. A few authors have acknowledged the confusion (e.g., Greenland and Robins 1986; Wickramaratne and Holford 1987; Weinberg 1993) and have suggested new ways of looking at the problem that might lead to more systematic analysis. Greenland and Robins (GR), in particular, have recognized the same basic principles and results that we have expounded here in Sections 6.2 and 6.3. Their analysis represents one of the few bright spots in the vast literature on confounding in that it treats confounding as an unknown causal quantity that is not directly measurable from observed data. They further acknowledge (as do Miettinen and Cook 1981) that the presence or absence of confounding should not be equated with absence or presence of collapsibility and that confounding should not be regarded as a parameter-dependent phenomenon.

However, the structural analysis presented in this chapter differs in a fundamental way from that of GR, who have pursued an approach based on judgment of "exchangeability." In Section 6.1 we encountered a related notion of exchangeability, one with which Lindley and Novick (1981) attempted to view Simpson's paradox; GR's idea of exchangeability is more concrete and more clearly applicable. Conceptually, the connection between confounding and exchangeability is as follows. If we undertake to assess the effect of some treatment, we ought to make sure that any response differences between the treated and the untreated group is due to the treatment itself and not to some intrinsic differences between the groups that are unrelated to the treatment. In other words, the two groups must resemble each other in all characteristics that have bearing on the response variable. In principle, we could have ended the definition of confounding at this point, declaring simply that the effect of treatment is unconfounded if the treated and untreated groups resemble each other in all relevant features. This definition, however, is too verbal in the sense that it is highly sensitive to interpretation of the terms "resemblance" and "relevance." To make it less informal, GR used De Finetti's twist of hypothetical permutation; instead of judging whether two groups are similar, the investigator is instructed to imagine a hypothetical *exchange* of the two groups (the treated group becomes untreated, and vice versa) and then to judge whether the observed data under the swap would be distinguishable from the actual data.

One can justifiably ask what has been gained by this mental exercise, relative to judging directly if the two groups are effectively identical. The gain is twofold. First, people are quite good at envisioning dynamic processes and can simulate the outcome of this swapping scenario from basic understanding of the processes that govern the response to treatment and the factors that affect the choice of treatment. Second, moving from

judgment about resemblance to judgment about probabilities permits us to cast those judgments in probabilistic notation and hence to invite the power and respectability of probability calculus.

Greenland and Robins made an important first step toward this formalization by bringing notation closer to where judgment originates – the human understanding of causal processes. The structural approach pursued in this book takes the next, natural step: formalizing the causal processes themselves.

Let A and B stand (respectively) for the treated and untreated groups, and let $P_{A1}(y)$ and $P_{A0}(y)$ stand (respectively) for the response distribution of group A under two hypothetical conditions, treatment and no treatment.[24] If our interest lies in some parameter μ of the response distribution, we designate by μ_{A1} and μ_{A0} the values of that parameter in the corresponding distribution $P_{A1}(y)$ and $P_{A0}(y)$, with μ_{B1} and μ_{B0} defined similarly for group B. In actuality, we measure the pair (μ_{A1}, μ_{B0}); after the hypothetical swap, we would measure (μ_{B1}, μ_{A0}). We define the groups to be *exchangeable* relative to parameter μ if the two pairs are indistinguishable, that is, if

$$(\mu_{A1}, \mu_{B0}) = (\mu_{B1}, \mu_{A0}).$$

In particular, if we define the causal effect by the difference $\text{CE} = \mu_{A1} - \mu_{A0}$, then exchangeability permits us to replace μ_{A0} with μ_{B0} and so obtain $\text{CE} = \mu_{A1} - \mu_{B0}$, which is measurable because both quantities are observed. Greenland and Robins thus declare the causal effect CE to be *unconfounded* if $\mu_{A0} = \mu_{B0}$.

If we compare this definition to that of (6.10), $P(y \mid do(x)) = P(y \mid x)$, we find that the two coincide if we rewrite the latter as $\mu[P(y \mid do(x))] = \mu[P(y \mid x)]$, where μ is the parameter of interest in the response distribution. However, the major difference between the structural and the GR approaches lies in the level of analysis. Structural modeling extends the formalization of confounding in two important directions. First, (6.10) is not submitted to direct human judgment but is derived mathematically from more elementary judgments concerning causal processes.[25] Second, the input judgments needed for the structural model are both qualitative and stable.

A simple example will illustrate the benefits of these features. Consider the following statement (Greenland 1998):

> (Q^*) "if the effect measure is the difference or ratio of response proportions, then the above phenomenon – noncollapsibility without confounding – cannot occur, nor can confounding occur without noncollapsibility." (pp. 905–6)

We have seen in this chapter that statement (Q^*) should be qualified in several ways and that, in general, noncollapsibility and confounding are two distinct notions – neither implying the other, regardless of the effect measure (Section 6.5.1). However, the

[24] In $do(\cdot)$ notation, we would write $P_{A1}(y) = P_A(y \mid do(X = 1))$.

[25] Recall that the $do(\cdot)$ operator is defined mathematically in terms of equation deletion in structural equation models; consequently, the verification of the nonconfounding condition $P(y \mid do(x)) = P(y \mid x)$ in a given model is not a matter of judgment but a subject of mathematical analysis.

question we wish to discuss here is methodological: What formalism would be appropriate for validating, refuting, or qualifying statements of this sort? Clearly, since (Q^*) makes a general claim about all instances, one counterexample would suffice to refute its general validity. But how do we construct such a counterexample? More generally, how do we construct examples that embody properties of confounding, effect bias, causal effects, experimental versus nonexperimental data, counterfactuals, and other causality-based concepts?

In probability theory, if we wish to refute a general statement about parameters and their relationship we need only present one density function f for which that relationship fails to hold. In propositional logic, in order to show that a sentence is false, we need only present one truth table T that satisfies the premises and violates the conclusions. What, then, is the mathematical object that should replace f or T when we wish to refute causal claims like statement (Q^*)? The corresponding object used in the exchangeability framework of Greenland and Robins is a counterfactual contingency table (see, e.g., Greenland et al. 1999b, p. 905, or Figure 1.7 in Section 1.4.4). For instance, to illustrate confounding, we need two such tables: one describing the hypothetical response of the treated group A to both treatment and nontreatment, and one describing the hypothetical response of the untreated group B to both treatment and nontreatment. If the tables show that the parameter μ_{A0}, computed from the hypothetical response of the treated group to no treatment, differs from μ_{B0}, computed from the actual response of the untreated group, then we have confounding on our hands.

Tables of this type can be constructed for simple problems involving one treatment and one response variable, but they become a nightmare when several covariates are involved or when we wish to impose certain constraints on those covariates. For example, we may wish to incorporate the standard assumption that a covariate Z does not lie on the causal pathway between treatment and response, or that Z has causal influence on Y, but such assumptions cannot conveniently be expressed in counterfactual contingency tables. As a result, the author of the claim to be refuted could always argue that the tables used in the counterexample may be inconsistent with the agreed assumptions.[26]

Such difficulties do not plague the structural representation of confounding. In this formalism, the appropriate object for exemplifying or refuting causal statements is a causal model, as defined in Chapter 3 and used throughout this book. Here, hypothetical responses (μ_{A0} and μ_{B0}) and contingency tables are not primitive quantities but rather are derivable from a set of equations that already embody the assumptions we wish to respect. Every parameterization of a structural model implies (using (3.51) or the $do(\cdot)$ operator) a specific set of counterfactual contingency tables that satisfies the input assumptions and exhibits the statistical properties displayed in the graph. For example, any parameterization of the graph in Figure 6.3 generates a set of counterfactual contingency tables that already embodies the assumptions that Z is not on the causal pathway between X and Y and that Z has no causal effect on Y, and almost every such parameterization will generate a counterexample to claim (Q^*). Moreover, we can also disprove (Q^*) by a casual inspection of the diagram and without generating numerical counterexamples.

[26] Readers who attempt to construct a counterexample to statement (Q^*) using counterfactual contingency tables will certainly appreciate this difficulty.

https://doi.org/10.1017/CBO9780511803161.008 Published online by Cambridge University Press

Figure 6.3, for example, shows vividly that the risk difference $P(y \mid x_1) - P(y \mid x_2)$ is not collapsible on Z and, simultaneously, that X and Y are (stably) unconfounded.

The difference between the two formulations is even more pronounced when we come to substantiate, not refute, generic claims about confounding. Here it is not enough to present a single contingency table; instead, we must demonstrate the validity of the claim for all tables that can possibly be constructed in compliance with the input assumptions. This task, as the reader surely realizes, is a hopeless exercise within the framework of contingency tables; it calls for a formalism in which assumptions can be stated succinctly and in which conclusions can be deduced by mathematical derivations. The structural semantics offers such formalism, as demonstrated by the many generic claims proven in this book (examples include Theorem 6.4.4 and Corollary 6.5.2).

As much as I admire the rigor introduced by Greenland and Robins's analysis through the framework of exchangeability, I am thoroughly convinced that the opacity and inflexibility of counterfactual contingency tables are largely responsible for the slow acceptance of the GR framework among epidemiologists and, as a by-product, for the lingering confusion that surrounds confounding in the statistical literature at large. I am likewise convinced that formulating claims and assumptions in the language of structural models will make the mathematical analysis of causation accessible to rank-and-file researchers and thus lead eventually to a total and natural disconfounding of confounding.

6.6 CONCLUSIONS

Past efforts to establish a theoretical connection between statistical associations (or collapsibility) and confounding have been unsuccessful for three reasons. First, the lack of mathematical language for expressing claims about causal relationships and effect bias has made it difficult to assess the disparity between the requirement of effect unbiasedness (Definition 6.2.1) and statistical criteria purporting to capture unbiasedness.[27] Second, the need to exclude barren proxies (Figure 6.3) from consideration has somehow escaped the attention of researchers. Finally, the distinction between stable and incidental unbiasedness has not received the attention it deserves and, as we observed in Example 6.3.3, no connection can be formed between associational criteria (or collapsibility) and confounding without a commitment to the notion of stability. Such commitment rests critically on the conception of a causal model as an assembly of autonomous mechanisms that may vary independently of one another (Aldrich 1989). It is only in anticipation of such independent variations that we are not content with incidental unbiasedness but rather seek conditions of stable unbiasedness. The mathematical formalization of this conception has led to related notions of *DAG-isomorph* (Pearl 1988b, p. 128), *stability*

[27] The majority of papers on collapsibility (e.g., Bishop 1971; Whittemore 1978; Wermuth 1987; Becher 1992; Geng 1992) motivate the topic by citing Simpson's paradox and the dangers of obtaining confounded effect estimates. Of these, only a handful pursue the study of confounding or effect estimates; most prefer to analyze the more manageable phenomenon of collapsibility as a stand-alone target. Some go as far as naming collapsibility "nonconfoundedness" (Grayson 1987; Steyer et al. 1997).

(Pearl and Verma 1991), and *faithfulness* (Spirtes et al. 1993), which assist in the elucidation of causal diagrams from sparse statistical associations (see Chapter 2). The same conception has evidently been shared by authors who aspired to connect associational criteria with confounding.

The advent of structural model analysis, assisted by graphical methods, offers a mathematical framework in which considerations of confounding can be formulated and managed more effectively. Using this framework, this chapter explicates the criterion of stable unbiasedness and shows that this criterion (i) has implicitly been the target of many investigations in epidemiology and biostatistics, and (ii) can be given operational statistical tests similar to those invoked in testing collapsibility. We further show (Section 6.5.3) that the structural framework overcomes basic cognitive and methodological barriers that have made confounding one of the most confused topics in the literature. It is therefore natural to predict that this framework will become the primary mathematical basis for future studies of confounding.

Acknowledgments

Discussions with James Robins and Sander Greenland were extremely valuable. Sander, in particular, gave many constructive comments on two early drafts and helped to keep them comprehensible to epidemiologists. Jan Koster called my attention to the connection between Stone's and Robins's criteria of no-confounding and caught several oversights in an earlier draft. Other helpful discussants were Michelle Pearl, Bill Shipley, Rolf Steyer, Stephen Stigler, and David Trichler.

Postscript for the Second Edition

Readers would be amused to learn that the first printing of this chapter did not stop statisticians' fascination with Simpson's nonparadox. Textbooks continue to marvel at the phenomenon (Moore and McCabe 2005), and researchers continue to chase its mathematical intricacies (Cox and Wermuth 2003) and visualizations (Rücker and Schumacher 2008) with the passion of the 1970–90s, without once mentioning the word "cause" and without once stopping to ask: "What's the point?" A notable exception is Larry Wasserman's *All Of Statistics* (Wasserman 2004), the first statistics textbook to treat Simpson's reversal in its correct causal context. My confidence in the eventual triumph of the causal understanding of this nonparadox has also been reinforced by a discussion published in the epidemiological literature, concluding in no ambiguous terms: "The explanations and solutions lie in causal reasoning which relies on background knowledge, not statistical criteria" (Arah 2008). It appears that, as we enter the age of causation, we should look to epidemiologists for guidance and wisdom.