6.1 Simpson's Paradox: An Anatomy

The reversal effect known as Simpson's paradox has briefly been 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

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

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

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

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

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

is not a statement about C being a positive causal factor for E, properly written

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

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 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

	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.

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 Eqs. (6.2)–(6.3)) and 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 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"

[1] The phrases "hold F fixed" or "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 to simulate "holding F fixed" by considering cases with equal values of F, namely, "conditioning" on F and $\neg F$, an operation 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).

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.

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

[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).

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 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 $P(E|C)$ with $P(C|E)$ 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. 41–42) 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 tormented 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 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 Fig. 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

[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).

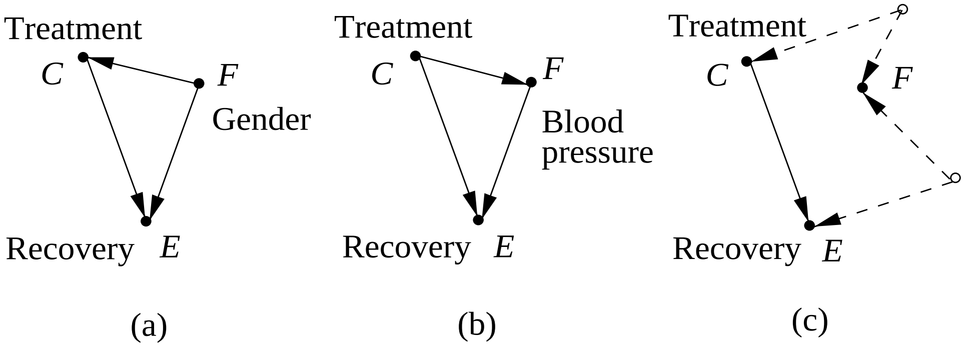

Figure 6.2: Three causal models capable of generating the data in Figure 6.1. Model (a) dictates use of the gender-specific tables, whereas (b) and (c) dictate use of the combined table.

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

[5] By "meta-statistical" I mean a criterion—not itself discernible—from statistical

by De Finetti (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 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 $F =$ female 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 con-

data for judging the adequacy of a certain statistical method.

template subjecting the new unit are equal to those prevailing while the group was observed. The reason we cannot use the combined table (Figure 6.1(a)) for determining the response of a new patient (with 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 $P(E|C)$; 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 only be understood 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-Rubins's 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.

[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 $P(F|C)$ instead of $P(F|do(C))$ and argued unconvincingly that we should equate $P(F|C) = P(F)$: "Instead [y]ou 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 $P(F|C) = P(F)$ in the example of Figure 6.2(b), we obtain the wrong result.

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. This 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 harmful effect on both males and females into formal statements in causal calculus:

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

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

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

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

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

[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 toange the classification of individual units as long as the relative sizes of the subpopulations remain unaltered.

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:

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

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

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

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

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

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

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

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, (6.8) cannot be derived and the difference $P(E|do(C)) - P(E|do(\neg C))$ may be positive or negative, depending on the relative magnitudes of $P(F|do(C))$ and $P(F|do(\neg C))$. 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., $P(E|do(C)) = P(E|C))$ or from a balanced study (i.e., $P(C|F) = P(C|\neg F))$. On the contrary; given the tables of Figure 6.1, our causal logic accepts gracefully that we are dealing with 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 of 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. Where 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.