Synthese
DOI 10.1007/s11229-010-9797-0

The logic of Simpson's paradox

Prasanta S. Bandyoapdhyay · Davin Nelson · Mark Greenwood · Gordon Brittan · Jesse Berwald

Received: 4 December 2009 / Revised: 3 May 2010 / Accepted: 26 July 2010

Abstract There are three distinct questions associated with Simpson's paradox. (i) Why or in what sense is Simpson's paradox a paradox? (ii) What is the proper analysis of the paradox? (iii) How one should proceed when confronted with a typical case of the paradox? We propose a "formal" answer to the first two questions which, among other things, includes deductive proofs for important theorems regarding Simpson's paradox. Our account contrasts sharply with Pearl's causal (and questionable) account of the first two questions. We argue that the "how to proceed question?" does not have a unique response, and that it depends on the context of the problem. We evaluate an objection to our account by comparing ours with Blyth's account of the paradox. Our research on the paradox suggests that the "how to proceed question" needs to be divorced from what makes Simpson's paradox "paradoxical."

Davin Nelson was a former student of Montana State University.

P. S. Bandyoapdhyay (✉) · G. Brittan
Department of History & Philosophy & Affiliate to Astrobiology Biogeocatalysis Research Center, Montana State University, Bozeman, MT, USA
e-mail: psb@montana.edu

G. Brittan
e-mail: uhigb@montana.edu

D. Nelson
Department of History & Philosophy, Montana State University, Bozeman, MT, USA

M. Greenwood · J. Berwald
Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA
e-mail: greenwood.stat@gmail.com

J. Berwald
e-mail: Berwald@math.montana.edu

Published online: 28 September 2010

Keywords Three questions · Conflation of three questions · Two experiments · Collapsiblity principle · Confounding · What to do questions

1 Overview

Simpson's Paradox (SP) involves the reversal of the direction of a comparison or the cessation of an association when data from several groups are combined to form a single whole. There are three distinct questions associated with SP: (i) why or in what sense, is SP a paradox? (ii) what is the proper analysis of this paradox? (iii) how should one proceed when confronted with a typical case of the paradox? We propose a "formal" account of the first two questions. We argue that there is no unique answer to the "how to proceed question?" Rather, what we should do varies as a function of the available background information.

One needs to be careful about the scope of the paper. We do not offer any novelty with regard to the treatment of the "how to proceed question" except by way of making clear the assumptions involved in addressing the latter and distinguishing the "how to proceed question" from what makes Simpson's paradox "paradoxical." Our analysis of the paradox, however, differs sharply from the causal account offered by Judea Pearl (Pearl 2000; Greenland et al. 1999).[1] In our view, his account is not persuasive. The premises that generate the paradox are non-causal in character and a genuine logical inconsistency is at stake when a full reconstruction of the paradox is carried out.

2 Simpson's paradox and its logical analysis

Consider an example of the paradox (hereafter called the type I version) (Table 1).

Here, "CV" includes two categorical variables, "F" for "females" and "M" for "men." "A" and "R" represent "the rates of acceptance/rejection" for two departments, D_1, and D_2. Here is a formulation of the paradox, in which the association in the sub-populations is reversed in the combined population. Although the acceptance rates for females are higher than for males in each department, in the combined population ignoring sex, the rates have reversed.

Table 1 Simpson's paradox (Type I)

CV	Dept. 1		Dept. 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	180	20	100	200	90	33	56
M	480	120	10	90	80	10	70

[1] The other influential work on causal inference is due to Spirtes, Glymour and Scheines and their colleagues (Spirtes et al. 2000). They are interested in representing systems of causal relationships as well as inferring causal relationships from purely observational data with the help of certain assumptions. How to address/eliminate situations like Simpsons's paradox in observational data while making causal inference is a key feature of their work. We have evaluated their research in Bandyopadhyay et al. (unpublished).

We now propose an analysis of the paradox. Consider two populations, [A, B] taken to be mutually exclusive and jointly exhaustive. The measured overall rates for each population are called, $[\alpha, \beta]$, respectively. Each population is partitioned into categories called, [1, 2], and the measured rates within each partition are called $[A_1, A_2, B_1, B_2]$. Let's assume that f_1 = the number of females accepted in D_1, F_1 = the total number of females applied to D_1; m_1 = the number of males accepted in D_1, M_1 = the total number of males applied to D_1. Then $A_1 = f_1/F_1$, and $B_1 = m_1/M_1$. Similarly, we could define A_2 and B_2. Let's assume that f_2 = the number of females accepted in D_2, F_2 = the total number of females applied to D_2; m_2 = the number of males accepted in D_2, and M_2 = the total number of males applied to D_2. So, $A_2 = f_2/F_2$ and $B_2 = m_2/M_2$. Likewise, we could understand α and β representing overall rates for each population, females and males, respectively. So the term $\alpha = \frac{(f_1+f_2)}{(F_1+F_2)}$ and $\beta = \frac{(m_1+m_2)}{(M_1+M_2)}$. Because $\alpha, \beta, A_1, A_2, B_1$ and B_2 are rates of some form, they will range between 0 and 1 inclusive. We stipulate the following definitions.

$$C_1 \equiv A_1 \geq B_1$$
$$C_2 \equiv A_2 \geq B_2$$
$$C_3 \equiv \beta \geq \alpha. \text{ We call } \mathbf{C} \equiv (C_1 \& C_2 \& C_3)$$

We define a term θ, which provides a connection between the acceptance rates (A_1, B_1, A_2 and B_2) within each partition to their overall acceptance rates (α and β).

$$\theta = (A_1 - B_1) + (A_2 - B_2) + (\beta - \alpha)\,.$$

A situation is a Simpson's paradox (SP) if and only if

(i) $\mathbf{C} \equiv (C_1 \& C_2 \& C_3)$ and
(ii) $\mathbf{C}_4 \equiv \theta = \{(A_1 - B_1) + (A_2 - B_2) + (\beta - \alpha)\} > 0$.

Each condition (i or ii) is necessary, but they jointly constitute sufficient conditions for generating SP. Consider why condition (i) alone is not sufficient. If **C** is true, then we get BOX I, because **C** generates the latter.

Box I

(1) $A_1 = B_1 \ \& \ A_2 = B_2 \ \& \ \beta = \alpha$;
(2) $A_1 = B_1 \ \& \ A_2 = B_2 \ \& \ \beta > \alpha$;
(3) $A_1 > B_1 \ \& \ A_2 = B_2 \ \& \ \beta = \alpha$;
(4) $A_1 = B_1 \ \& \ A_2 > B_2 \ \& \ \beta = \alpha$;
(5) $A_1 > B_1 \ \& \ A_2 > B_2 \ \& \ \beta = \alpha$;
(6) $A_1 > B_1 \ \& \ A_2 = B_2 \ \& \ \beta > \alpha$;
(7) $A_1 = B_1 \ \& \ A_2 > B_2 \ \& \ \beta > \alpha$; and finally
(8) $A_1 > B_1 \ \& \ A_2 > B_2 \ \& \ \beta > \alpha$.

Case 1, i.e., $A_1 = B_1 \ \& \ A_2 = B_2 \ \& \ \beta = \alpha$ (in Box I) shows that the condition (i) (i.e., **C**) taken alone is not sufficient because the case 1 implies neither the cessation

Table 2 No Simpson's paradox

CV	Dept. 1		Dept. 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	40	60	100	100	40	50	46.6
M	50	50	80	120	50	40	43.3

of the association nor reversal in the overall population. Therefore, we need (ii) $\theta > 0$ to eliminate case 1. Hence, **C** can't alone be sufficient. We provide an example using Table 2 to argue why (ii) alone is not sufficient.

In Table 2, $A_1 > B_1$, $B_2 > A_2$ and $\beta > \alpha$. This example satisfies (ii) because it implies θ is greater than 0. However, this is not a case of Simpson's paradox. It has violated **C** because one condition of **C**, $A_2 \geq B_2$, remains unsatisfied. Hence, (ii) cannot be solely adequate to generate Simpson's paradox. If we have both **C** and C_4 (i.e., $\theta > 0$) then we have the sufficient condition for generating the paradox.

Consider why (i) is necessary. To answer this we need to show that if **C** is not satisfied, then we won't be able to derive the paradox. If we deny **C**, then we get seven combinations (Box II).

Box II

1. $\sim C_1$
2. $\sim C_2$
3. $\sim C_3$
4. $\sim C_1$ &$\sim C_2$
5. $\sim C_2$ & $\sim C_3$
6. $\sim C_1$ & $\sim C_3$
7. $\sim C_1$ & $\sim C_2$ & $\sim C_3$

We will show that if $\sim C_1$ is the case (i.e., case 1 in Box II), then we get this combination where $B_1 > A_1$, $A_2 \geq B_2$ and $\beta \geq \alpha$. This does not manifest any reversal, hence can't be a case of Simpson's paradox as shown in Table 3.

Table 3 No Simpson's paradox

CV	Dept. 1		Dept. 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	70	30	30	70	70	30	50
M	40	60	50	50	40	50	45

Given BOX II if we take other cases in which **C** is false, it will also follow that there will be no paradox. Since proofs for the negations are similar to the one just given for $\sim C_1$, we do not repeat them. Hence, **C** is necessary.

Why is (ii), that is, $C_4 : \theta > 0$, necessary? If **C** is true, then $\theta \geq 0$. If **C** is true, but $\theta > 0$ is not necessary, then the denial of $\theta > 0$ implies $\theta \leq 0$. The latter, $\theta \leq 0$, implies that disjunction (i.e., $\theta < 0$ or $\theta = 0$) is true. If either one disjunct (i.e., $\theta = 0$ or $\theta < 0$) is true, then it implies that the entire disjunction, $\theta \leq 0$, is true. If $\theta = 0$ is true then it entails case 1, (i.e., $A_1 = B_1 \& A_2 = B_2 \& \beta = \alpha$) of BOX I, which does not represent an instance of reversal. Hence, (ii) $\theta > 0$ must be necessary.

There are two points worth-mentioning. First, the characterization of the puzzle in terms of our two conditions captures the central intuitions at stake in the example given; they are in no way ad hoc. The central intuitions are, once again, the reversal and the cessation of an association in the overall population. Three more points follow. First, the paradox is "structural" in character, in the sense that the reasoning that leads to it is deductive. Consider our example, which involves simple arithmetic. The overall rates of acceptance for both females and males follow from their rates of acceptance in two departments taken separately. Second, a probabilistic (in the sense of a statistical or inductive) solution is not available.[2] Third, unless someone uses the notion of causation trivially, for example, believes that 2+2 "causes" 4, there is no reason to assume that there are causal intuitions lurking in the background.[3]

So far we have discussed SP in general. We are now interested in knowing specific relationships between two rates of acceptance in each sub-population for both populations. We have proved two theorems, which we call Theorems 1 and 2, to address these relationships. We have also proved a derived result from Theorems 1 and 2, which we call Theorem 3, showing the connection between them. In addition, once we know the interrelationships between two rates of acceptance in each sub-population, we want to know whether we have logical relationships between those rates of acceptance in each sub-population to their overall acceptance rate in each population. A set of lemmas satisfy our curiosity by proving those relationships. First, we

[2] This account seems to go against the view held by Freedman et al. (1999). However, this depends on how we construe the following passage along with an email communication with David Freedman. These authors write in their celebrated textbook, "[t]he statistical lesson: relationships between percentages in subgroups (for instance, admissions rates for men and women in each department separately) can be reversed when the subgroup are combined. This is called *Simpsons's paradox*." If we construe "statistical lesson" in terms of non-monotonic reasoning, then it does not seem that there is any statistical lesson hidden in the paradox. However, it is possible that in this passage the mathematical reasoning involved in the paradox has been taken broadly to stand for statistical reasoning. If the second construal is its intended meaning, then there is no difference between their argument and ours. The latter meaning is what has been hinted by Freedman in his response to one of the author's query. Freedman wrote, "[t]he issue [concerning the paradox] is not uncertainty. It has little to do with the distinction between inductive and deductive reasoning, as far as I understand these terms. Simpsons's paradox is a surprising fact about weighted averages, i.e., it's a math fact. It has big implications for applied statistics (18 January, 2004)." We agree with Freedman that it has nothing to do with uncertainty and is a mathematical fact about ratios, but we disagree with him about the nature of reasoning which, according to us, is purely deductive. Readers are invited to compare this footnote and our comments on Freedman's email with the footnotes 12 and 13 and our comments in the corresponding body of the paper. We are very much thankful to Freedman for this communication.

[3] This goes against the view held by Pearl (2000, 2009).

Table 4 No Simpson's paradox

CV	Dept 1		Dept 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	75	225	75	225	25	25	25
M	10	90	20	80	10	20	15

Table 5 No Simpson's paradox

CV	Dept 1		Dept 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	10	90	20	80	10	20	15
M	75	225	75	225	25	25	25

motivate each theorem along with lemmas using examples; their proofs are provided in the appendix. Theorems 1, 2, and 3 are given below:

TH1: SP results only if $A_1 \neq A_2$.
TH2: SP arises only if $B_1 \neq B_2$.
TH3: SP arises only if $(A_1 \neq A_2)$ if and only if $(B_1 \neq B_2)$.

The following example based on Table 4 shows why the condition for Theorem 1 needs to hold.

Since $A_1 = A_2$, i.e., 25% = 25%, no paradox results. Table 5 explains why the condition laid down in Theorem 2 needs to hold.

Since $B_1 = B_2$, i.e., 25% = 25%, a paradox does not result in Table 5. The first two theorems provide us with the information that neither $A_1 = A_2$ nor $B_1 = B_2$ can hold in cases of SP. We can see from our examples, based on Tables 4 and 5, that there might be some relationship between $A_1 \neq A_2$ and $B_1 \neq B_2$. However, those examples can't tell us exactly what those relationships are. Theorem 3 has convincingly showed that relationship, which is, Simpson's paradox arises only if $(A_1 \neq A_2)$ if and only if $(B_1 \neq B_2)$. Given our characterization of Simpson's paradox, we realize that α being the overall rate of acceptance for population A, it is a weighted average of A_1 and A_2. Hence, α lies in between A_1 and A_2. Similarly, β being a weighted average of B_1 and B_2, β also lies in between B_1 and B_2. Our four lemmas (LM1, LM2, LM3, and LM4) will provide us with more specific information between the relationships of different rates of acceptance in the sub-populations and the overall rate of acceptance in each population. They tell us what specific bridge we are able to build between the rates of acceptance, e.g., $A_1 \neq A_2$, or $B_1 \neq B_2$ in each subpopulation and their overall acceptance rates in each population.

We have two lemmas (LM1 and LM2) showing the inter-connections among A_1, A_2, and α. Lemma 1 and lemma 2 are given below:

LM1: If $A_1 > A_2$, then $A_1 > \alpha > A_2$.
LM2: If $A_2 > A_1$, then $A_2 > \alpha > A_1$.

Table 6 Simpson's paradox

CV	Dept. 1		Dept. 2		Acceptance rates		Overall acceptance rates (%)
	A	R	A	R	D_1 (%)	D_2 (%)	
M	25	75	100	100	25	50	42
F	75	225	100	100	25	50	35

Table 1 type I of Simpson's paradox obeys LM1 in which when $A_1 = 90\% > A_2 = 33\%$, $A_1 = 90\% > \alpha = 56\% > A_2 = 33\%$. Table 6 obeys LM2. Here, when $A_2 = 50\% > A_1 = 25\%$, $A_2 = 50\% > \alpha = 42\% > A_1 = 25\%$. Proof for LM1 has been furnished in the appendix.

LM3 and LM4 are provided below:

LM3: If $B_1 > B_2$, then $B_1 > \beta > B_2$.
LM4: If $B_2 > B_1$, then $B_2 > \beta > B_1$.

Simpson's paradox type I gives an example in which the condition for LM3 holds. In the type I version, when $B_1 = 80\% > B_2 = 10\%$, we find $B_1 = 80\% > \beta = 70\% > B_2 = 10\%$. Likewise, the following table satisfies the condition for LM4.

Table 6 shows that when $B_2 = 50\% > B_1 = 25\%$, $B_2 = 50\% > \beta = 35\% > B_1 = 25\%$. These relationships among four lemmas are symmetric with respect to the indices. Therefore, we have proved only LM1 and LM3. The other cases, LM2 and LM4, are handled identically by swapping variables and indices.

This straightforward "formal" analysis might not, however, alleviate the suspicions of those who are not familiar with the literature but who, when confronted with SP examples, find them "perplexing" and want a "deeper" explanation of their puzzlement. We provide an explanation of how the paradox arises in our type I version and why people find it perplexing. To explain each, we have reconstructed our type I version of SP in terms of its premises and conclusion. However, the point of the reconstruction is adequately general to be applicable to all types of SP. Before the reconstruction, we introduce a principle called the collapsibility principle (CP) which plays a crucial role in the reconstruction. We call a dataset collapsible if and only if $[A_1 \geq B_1$ and $A_2 \geq B_2] \rightarrow \alpha \geq \beta$. We call the principle that underlies this dataset, the CP. Recall, A_1 and A_2 stand for the rates of acceptance for population A in departments 1 and 2, respectively. Similarly, B_1 and B_2 stand for the rates of acceptance for population B in departments 1 and 2, respectively. In contrast, α and β are rates of acceptance for A and B populations in the overall school. More explicitly, if we use our earlier notations of f_1, F_2, m_1, M_2, then the CP implies $[(f_1/F_1) > (m_1/M_2)\&(f_2/F_2) > (m_2/M_2)] \rightarrow ((f_1 + f_2)/(F_1 + F_2)) > ((m_1 + m_2)/(M_1 + M_2))$. In the type I version, even though the data set satisfies the antecedent, that is, A_1 (i.e., f_1/F_2) $> B_1$ (i.e., m_1/M_1) and A_2 (i.e., f_2/F_2) $> B_2$ (i.e., m_2/M_2), its consequent remains unsatisfied, that is α, i.e., $((f_1 + f_2)/(F_1 + F_2)) < \beta$, i.e., $((m_1 + m_2)/(M_1 + M_2))$. Here is the reconstruction of the type I version of SP.[4]

[4] We by and large agree with those who think that SP is not really a paradox. That is, the reversal or cessation follows from arithmetic premises. However, it does not explain away why most people not previously

P1: Female and male populations are mutually exclusive and jointly exhaustive and one can't be a student of both departments along with satisfying two conditions (i & ii) in our characterization of what is called SP.
P2: The acceptance rate of females is higher than that of males in department # 1.
P3: The acceptance rate of females is higher than that of males in department # 2.
P4: If P2 & P3 are true, then the acceptance rate for females is higher than that of males overall.
P5: However, fewer females are admitted overall. (That is, the consequent of P4 becomes false.)

Conclusion: the deductive consequence of P2, P3, P4 and P5 contradict one another. There is a genuine paradox involved.

In our derivation of the paradox, premise 4 plays a crucial role. It rests on the CP. In our type I version, the rates of acceptance for females are greater than those of males in each department. That is, $A_1 > B_1$ and $A_2 > B_2$, but $\alpha < \beta$. In fact, that the CP is not generally true is shown by our derivation of a contradiction.

So our answer to the question, why do so many individuals find the paradox startling? is simply that humans tend to invoke the CP uncritically, as a rule of thumb, and thereby make mistakes in certain cases about proportions and ratios;[5] they find it paradoxical when their usual expectation, i.e., the CP is applicable across the board, captured in premise 4, turns out to be incorrect.[6] There is, however, an alternative account of the paradox. The fact that it is so well-known justifies its brief consideration.

3 Pearl's causal account of the paradox

Pearl argues that the correct diagnosis of the paradox lies in understanding it in causal terms. In his view, the arithmetical conclusions reached seem counter-intuitive only because we commonly make two incompatible assumptions, that causal relationships are governed by the laws of probability and that there are certain (non-probabilistic) causal assumptions we share among ourselves about the world. The operative causal assumption to which he refers is that where there is correlation, there must be an underlying cause. The source of our perplexity here is that there cannot be a cause that would simultaneously account for incompatible correlations, the lower and higher

Footnote 4 continued
acquainted with SP find it perplexing. This logical reconstruction pinpoints that perplexing premise in the logical reconstruction of the paradox.

[5] Our empirical research on students over the last 4 years has vindicated our claim. Interestingly, in the last year 83% of 106 responses consistently committed the same type of mistake in the story-driven SP type situation whereas 57% of them committed the error in the formula-driven SP situation (See Sect. 7 for more on our experiments on SP). Although it is an empirical finding, it could be explained within our analysis of the paradox. Details of the protocol for the experiment could be provided on request. Caleb Galloway first suggested to us the idea of running the experiments on the paradox.

[6] One might be tempted to contend that the invocation of the CP could cut in either way, a causal or non-causal way. According to this objection, the reader is yet to be convinced that the CP is a non-causal principle. In Sects. 3 and 7, we have addressed this point and argued that it is entirely a non-causal principle which underlies all versions of the paradox.

rates of male/female acceptance.[7] Once we reject either of these assumptions, and he opts for rejecting the first, the "paradox" is no longer paradoxical. On the other hand, when we don't distinguish causal from statistical hypotheses, we are confronted with the paradox.

Pearl's resolution of the paradox emerges from his general approach to causal hypotheses as distinguished from statistical hypotheses. He makes two basic points. One is that SP arises from mixing explicit statistical and implicit causal considerations together. The notion he uses to explain the paradox is "confounding" which, he argues, is a causal notion. In the type I version, for example, the effect on acceptance (A) of the explanatory variable, sex, (S) is hopelessly mixed up (or "confounded") with the effects on A of the other variable, department (D). According to him, we are interested in the direct effect of sex on acceptance and not an indirect effect by way of another variable such as department. The effect of S on A is confounded with the effect on A of a third variable D.

Pearl's other point is that causal relationships are more *stable* than statistical relationships and therefore, causal hypotheses often cannot be analyzed in statistical terms (Pearl 2000, p. 25, 2009, p. 25). Suppose we would like to know Bill Clinton's place in US history had he not met Monica Lewinsky (Pearl 2000, p. 34, 2009, p. 34). Most people now agree that it would be very different. However, there is no statistical model one could construct that would provide the joint occurrence of Clinton and no Lewinsky. There simply are no appropriate data, as there are, for instance, in a fair coin-flipping experiment. As in the case of the latter in which we have a good understanding about the joint probability of two fair coins, we lack such an understanding in the case of the former because we don't know the joint probability of Clinton and no Lewinsky.[8]

In his paper with Greenland and Robins, Pearl thinks confounding is sometimes confused with a non-causal notion, non-collapsibility. In fact, in any version of SP, the data set will be non-collapsible. For example, as we have argued before, $A_1 > B_1$ and $A_2 > B_2$, but $\alpha < \beta$. This leads some to conclude that non-collapsibility is synonymous with confounding. Although in SP examples, it seems that non-collapsibility and confounding go hand in hand, Pearl thinks rightly that they are conceptually different notions (Pearl 2000, p. 193). A simple example should make this clear.

Assume that we have observed that clouds are often followed by a good crop. Based on this statistical information, we could make a causal inference connecting "X", which stands for clouds, to "Y," which stands for good crops. It could be the case that there

[7] See Pearl (2000, p. 180) and (2009) with the same page number. However, he has here used the example of the effect of drugs on males and females.

[8] One could provide a response to Pearl by contending that we (might) have data for the US presidents being womanizers as well as how much they are respected in the US presidential history. According to this response, based on these data, we might be able to retrieve a statistic concerning how much Bill Clinton would be remembered even with Lewinsky in the later day presidential history without deploying any counterfactual strategy at the core of a causal account of the paradox. One seeming drawback with this response is that the reference class problem for the frequency interpretation might also arise here regarding whether the data for the US presidents, being womanizers as well as how much they are revered, could be applied to a single case event like "Clinton and Lewinsky". We thank Elliott Sober for calling our attention to this way of thinking about this issue.

is another variable "Z," which is in fact highly correlated with X. "Z" is the variable that farmers plough their land when there are clouds in the sky. If Z is unaccounted for in the model when the effect of X on Y is made, the effect of X that appears could be due to the influence of Z. Under this scenario, Z is said to have confounded the effect of X. Given our characterization of collapsibility, we don't know whether the data set is collapsible, although we can tell for sure that there is confounding. Pearl insists that without a proper model specification—one where possible confounding facts are accounted for—it is not possible to parcel out the unique effects of X on Z. In other words, we need to either eliminate Z as a causal mechanism or more accurately estimate the effect it (and X) have on Y.

4 An evaluation of Pearl's account

Pearl's analysis is ingenious. But that SP need not rest on mixing causal and statistical considerations follows at once from the fact that our derivation of it involves neither. It is not easy to come up with an example which precludes invoking some sort of appeal to "causal intuitions." But what follows is, we think, such a case.[9] It tests in a crucial way the persuasiveness of Pearl's account.

Suppose we have two bags of marbles, all of which are either big or small, and red or blue. Suppose in each bag, the proportion of big marbles that are red is greater than the portion of small marbles that are red. Now suppose we pour all the marbles from both bags into a box. Would we expect the portion of big marbles in the box that are red to be greater than the portion of small marbles in the box that are red? Most of us would be surprised to find that our usual expectation is incorrect. The big balls in bag 1 have a higher ratio of red to blue balls than do the small balls; the same is true about the ratio in bag 2. But considering all the balls together, the small balls have a higher ratio of reds to blues than the big balls do. We argue that this is a case of SP since it has the same mathematical structure as the type I version of Simpson's paradox. There are no causal assumptions made in this case, no possible "confounding." But it still seems surprising.[10] That is the point of the test case. The proponents of a causal analysis of the paradox must argue either that this is not surprising or that it engages in causal reasoning even when the question presents us with nothing causal. We find neither of these replies is tenable. We believe the test case shows that at least sometimes there is a purely mathematical mistake about ratios that people customarily make.

[9] This counterexample has been suggested to us by John G. Bennett.

[10] One might wonder whether Pearl could maintain his causal stance toward the paradox while conceding this case as a case of Simpson's paradox which is non-causal. In fact, we don't think that Pearl could adopt this (weaker) position. First, his book (Pearl 2000, 2009) and papers on this issue do not allow any such endorsement. Second, if he were to adopt this weaker position about the paradox, then this would imply that there are at least two types of SP. One is non-causal and the second one is causal. As it were, we advocate the former and Pearl the latter giving the impression that both positions could nicely co-exist with regard to the paradox. In fact, contra this position, we argue that there is only one type of Simpson's paradox which is non-causal with which both its paradoxical nature and the account to provide a correct analysis of the paradox could be explained away in terms of mathematical notions like ratios/proportions which are non-causal. (For more on this point, see the end of this section along with Sect. 7.)

It must be admitted that there are all sorts of complexities about going from correlation to causation. Correlations are not causes, though correlations are (part of the) evidence for causes. But what is paradoxical in the SP case has little to do with these complexities; there is simply a mistaken inference about correlations, which are really just ratios. Of course, when there are different correlations available which may seem to support conflicting causal inferences, the inference from correlations to cause becomes much more difficult; no one could reasonably deny that. But the paradoxical nature of the examples really lies in the part that involves the mistaken assumptions about the correlations (ratios) themselves.

In our reconstruction of the paradox, we suggested that human beings are not good at reasoning concerning ratios. We are not the first, of course, to point out that human beings are not very good at these sorts of computations. What we have done is to isolate the mistaken assumption usually made and to provide empirical support for our claim.[11] Pearl himself talks about mistaken numerical assumptions, but proceeds at once to interject causal considerations. He writes,

> 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 mind governed by the calculus of proportions, Fig. 6.3 [i.e., an example of Simpson's paradox] would have evoked no surprise at all and Simpson's paradox would never have generated the attention that it did. (*Causality*, 2000, p. 182 and 2009, p. 182.).

Pearl's point may be reconstructed as follows.

> Human beings are not good at (transitory) ratios and proportions. To remedy this defect, they import (invariant) causal notions, in the process confusing collapsibility with confounding. If there were no confusion between the two, there would be no paradox (or rather, there would be no "perplexity.")

Although one can sympathize with the claim that humans often tend to see causes where they should not, it is enough here to point out, once again, that mistaken numerical assumptions suffice to demonstrate the paradox; jumping to conclusions does not necessarily require that we are pushed by our causal intuitions. We certainly admit that surprising facts about proportions come up frequently when we infer causes from proportions. This is when our mistakes about proportions seem most troubling to us. In this respect, the test case we contrived is rather unusual. But it proves our point.

5 How to proceed questions

In the case of SP, "how to proceed?" questions arise when investigators are confronted with choosing between two conflicting statistics, for example, in Table 1, (i) the uncombined two departments' statistics and (ii) their combined statistics. Which

[11] See footnote 5 for the confirmation of this point.

Table 7 Simpson's paradox (medical example)

CV	M		∼ M		Recovery rates		Overall recovery rates (%)
	R	∼ R	R	∼ R	M (%)	∼ M (%)	
T	18	12	2	8	60	20	50
∼ T	7	3	9	21	70	30	40

Table 8 Simpson's paradox (agricultural example)

CV	T		∼ T		Yield rates		Overall yield rates (%)
	Y	∼ Y	Y	∼ Y	T (%)	∼ T (%)	
W	18	12	2	8	60	20	50
∼ W	7	3	9	21	70	30	40

one should they use to recommend action? Our reply is that there is no unique response to *all* versions of SP; the response depends on the specific nature of the problem at issue. However, one could provide stable recommendations for certain versions of the paradox when we assume certain features about these versions to be correct over and above the data at hand.

One needs to be circumspect that not *all* versions of SP necessarily involve the "how to proceed?" questions. The test case, for example, asks "would we expect the proportion of big marbles in the box that are red to be greater than the proportion of small marbles in the box that are not red?" No straightforward question concerning whether to recommend an action from sub-groups or whole seems to be at stake. However, it is evident that many interesting cases of recommending actions arise when we infer causes/patterns from proportions. The standard examples (Lindley and Novick 1981; Meek and Glymour 1994; Pearl 2000, 2009) deal with cases in which "how to proceed" questions become preeminent. But it should be clear in what follows that there is no unique response to this sort of question for all varieties of the paradox.

Consider Table 7 based on data about 80 patients. 40 patients were given the treatment, T, and 40 assigned to a control, ∼ T. Patients either recovered, R, or didn't recover, ∼ R. There were two types of patients, (i) males (M) and (ii) females (∼M).

One would think that treatment is preferable to control in the combined statistics, whereas, given the statistics of the sub-population, one gathers the impression that control is better for both men and women. Given a person of unknown sex, would one recommend the control? The standard response is clear. That is, the control ($Pr(R|\sim T) > Pr(R|T)$) is better for a person of unknown sex. Call this first example the medical example. In the second example, call it the agricultural example, we are asked to consider the same data, but now we are asked to replace T and ∼ T by the varieties of plants (white [W] or black variety [∼ W]), R and ∼ R by the yield (high[Y] or low yield [∼Y]) and M and ∼M by tall and short plants ([T] or [∼ T]) (Table 8).

Given this new interpretation, the overall yield rate suggests that planting the white variety is preferable since it is 10% better overall, although the white variety is 10%

worse among both tall and short plants (sub-population statistics). Which statistics should one follow in choosing between which varieties to plant in the future? The standard recommendation is that in this case one should take the combined statistics and thus recommend the white variety for planting, ($Pr(Y|W) > Pr(\sim Y|W)$), which is in stark contrast with the recommendation given in the medical case. In short, both medical and agricultural examples provide varying responses to the "how to proceed question?" There is no unique response regarding which statistics, subpopulation or whole, to follow in every case of SP. We agree with standard recommendations with a proviso, i.e., we need to use substantial background information to answer how to proceed questions. These recommendations are standard because they are agreed upon by philosophers (e.g., Meek and Glymour), statisticians (e.g., Lindley and Novick), and computer scientists (e.g, Pearl). To explain which assumption(s) is at stake in these two examples, we will confine ourselves to Pearl's causal analyses of these examples. According to Pearl, those background assumptions are primarily causal assumptions that go beyond the data presented in the two tables as well as those assumptions that underlie probabilistic calculations capturing our preferences for one course of action over the other.

Consider his view with regard to the medical example in which we are confronted with the question "should treatment or control be recommended to a patient of the unknown sex?" We think that there are three assumptions at work in his analysis. Two of them are causal and one is an ethical one. The first causal assumption is that the unknown sex to which the treatment or control will be administered is contingent on whether this new individual shares the same causal conditions with the group we have studied. Whether the conditions for the two groups, the group we have studied with the help of the tables and the group from which the individual of that unknown sex comes, are the same or whether probabilities will remain invariant across the groups depends on making causal assumptions. He also thinks that there might be a difference between these two groups in terms of their causal conditions. That is, the group studied made their decisions whether to receive treatment or control. In contrast, the individual of an unknown sex will be given the treatment or control randomly without deliberate choice.

At any rate, his rationale for recommending control to the individual of an unknown sex also has an ethical dimension to it. Since whether to recommend treatment or control to the unknown sex depends on substantial causal assumptions about the population from which that individual has been taken and which assumptions are operative for that individual are unknown, it is, therefore, *safer* to recommend control to that individual. In addition, there is another causal assumption behind his analysis. There is significant confounding going on among three variables, "treatment", "sex/gender", and "recovery." Interestingly, Pearl is mostly interested in this casual consideration of confounding while recommending "control" in the medical example. In this regard, the effect of treatment on recovery is confounded with the effect of sex on recovery. We are interested in knowing whether to recommend treatment or control to a subject of unknown sex. Hoping that the combined statistics would provide us with the required information, we looked at the combined table. However, since there is a significant effect of confounding in the combined table, the sub-population statistics are the right statistics to look for and based on those statistics we should recommend

control, because the sub-population shows clearly the confounding effect of gender on the recovery. Therefore, in the case of the medical example, recommending control is based on taking the sub-population statistics as the guide to our action.

Pearl argues, however, that in the case of the agricultural example, the operative causal conditions are vastly different from the medical example. No rational deliberation regarding whether to receive treatment is present in the case of the agricultural example. In addition, there is no significant confounding going on. In fact, he could offer this causal account about the agricultural example. Both yield and height are consequences of the variety. The white variety causes plants to grow tall which in turn causes high yield. Being tall increases the chance of high yield and is correlated with exposure (i.e., white variety), since in the latter case, being tall could be a result of the white variety. It could even be a causal factor for high yield among non-exposed plants, because high yield could result from a cause other than being white. For example, taller plants might easily receive more sun-light than shorter plants, leading to high yield from the former. Being tall can't be considered a purely confounding factor, since the effect of being white is mediated through the effect of tall plants. Any factor that represents a step in the causal chain between the exposure (white variety) and disease (yield) should not be treated as an extraneous confounding factor, but instead requires special treatment as an intermediate factor.

In the agricultural example, we were interested in knowing whether to recommend white or black variety for future planting. Hoping that the combined table will give us the correct guidelines, we looked at the latter. Since there is no significant confounding effect of the length of the plant on the variety, we use the combined statistics to support our decision that one should plant the white variety. Although there might be a normative element embedded in any such recommendation (because a wrong decision could result in economic losses in terms of bad crops), the magnitude of this normative consideration is far outweighed by the mere ethical considerations in the medical example.

Three points need to be mentioned. First, there is no point in denying that there are causal considerations involved in both examples, the medicine and agricultural examples. They have no doubt contributed to our understanding regarding how to address the "what to do question"? Second, the first point, however, does not imply that the notion of utility is irrelevant to this question. In the medical example, the utility of recommending "control" to a person of unknown gender has to be taken into consideration. What if that particular individual with unknown gender does have certain physiological issues which might react badly with "treatment"? In this case, we will be making a wrong decision if we recommend "treatment" to her. So we need to take into account the disutility associated with this possible scenario. Likewise for the agricultural example, we might run some amount of risk in recommending "white variety" as stated above. None the less, making a wrong decision is not as terrible in the agricultural example as in the case of medicine example. Although it is evident what sorts of utilities are involved in addressing "what to do questions" insofar as both SP type situations are concerned and these are likely reasons why we don't bring them to the table when confronted with "what to do questions", we can't afford to overlook those implicit considerations of utilities in making a decision. The third point has to do with the distinction between what makes SP paradoxical and what to

Table 9 A comparison between two sets of conditions

Blyth's conditions	Our Conditions
$C_1 : A_1 \geq B_1$	$C_1 : A_1 \geq B_1$
$C_2 : A_2 \geq B_2$	$C_2 : A_2 \geq B_2$
$C_3' : \beta > \alpha$	$C_3 : \beta \geq \alpha$
	$C_4 : \theta > 0$

do when confronted with SP type situations or how to infer a cause from a correlation. In our account, we have primarily addressed two questions (i) why SP is paradoxical? And (ii) what is the correct analysis of the paradox? Providing an explanation for what makes SP paradoxical does not, however, provide information specific to each version of the paradox. Even with the same data, two different sets of assumptions have led to different recommendations regarding the "what to do question." Neither does providing a correct analysis of the paradox entitle one to address the "what to do question". Because our account is mainly concerned with the mathematical structure of SP and how the latter provides its correct analysis, it does not directly tell us anything significant about the "what to do question" in contrast to Pearl and other causal theorists who offer illuminating recommendations for this question. To repeat ourselves, Pearl's account gains plausibility by blurring the difference between our three questions, loses plausibility once they are distinguished as we have here.

6 A comparison with Blyth's account of the paradox

It is alleged that there is a striking similarity between Blyth (1972) celebrated paper on SP and our account as both are formally motivated beginning with some initial conditions of SP and then define it in terms of those conditions.[12] Consequently, according to this objection, although our account is correct, it, however, fails to deliver any new information about the paradox than what has already been contained in Blyth's paper.[13] To address this charge, we will first discuss Blyth's treatment of SP and then evaluate the charge based on several grounds. To make an easy transition between our notation and his, we write his conditions in terms of our notation shown in Table 9.

As one could see, Blyth's first two conditions are the same as ours. But, his third condition is expressed in terms of a strict inequality, whereas ours is a weaker condition. Most importantly, his conditions imply two features which deserve mention. First, he does not allow the cessation of associations between variables in the overall population, when there is a strict inequality in subpopulations, as a case of SP, although ours do. From our first three conditions, it follows, we could have a case of SP where we have $A_1 > B_1 \& A_2 > B_2$, but $\beta = \alpha$ (see Table 10 below in this section for this possibility). Second, from our three conditions, we are able to derive

[12] One of the referees of a leading journal has raised this objection.

[13] To be faithful to this objection, the referee admits that the only difference between ours and Blyth's is that we have provided two experiments about the paradox which, according to the referee, is the only novelty of the paper (see Sect. 7 for those two experiments).

Table 10 Simpson's paradox

CV	Dept. 1		Dept. 2		Acceptance rates		Overall acceptance rates (%)
	Accept	Reject	Accept	Reject	Dept. 1 (%)	Dept. 2 (%)	
F	90	1410	110	390	6	22	10
M	20	980	380	2620	2	12	10
				Diff: F-M	4	10	0

a case in which $A_1 = B_1 \& A_2 = B_2$ as well as $\beta = \alpha$, which is clearly not a case of SP. To eliminate this possibility, we have introduced C_4 which states that $\theta > 0$, when θ is defined as $(A_1 - B_1) + (A_2 - B_2) + (\beta - \alpha)$. However, he does not need a condition like C_4 because his third condition will automatically prevent this case (i.e., $A_1 = B_1 \& A_2 = B_2$ as well as $\beta = \alpha$) from occurring.

Blyth prefers to construe SP in terms of interaction effects of two variables. An interaction effect is the one in which combined effects of two variables is not a simple sum of their separate effects. Since one purpose of our account is to provide an analysis of the possible ways SP might be generated, the theme of interaction effect does not directly have any connection with how SP results from interaction effects of two variables. This is one fundamental difference between his and ours. Since this section is intended to compare his account with ours, we will adopt his interaction locutions to see what follows from his argument. Blyth writes, "[t]he paradox (i.e., SP) can be said to result from... interaction of B and C." It can't be the case that he takes "interaction effect" as a sufficient condition for the emergence of SP. In Table 2, we came across "interactions" between "gender" and "departments"; yet those interactions do not result in SP. Therefore, interaction effects can't be sufficient for generating SP. The interaction effect means that the effect of one variable is different depending on which group of the other variable one is considering. For this example, it means that the difference in the acceptance rates between the sexes is different between the departments. This does not, however, address whether there is a reversal or cessation which, as we know, is at the core of the paradox.

In contrast, if we take "interaction effect" as necessary, and incidentally, this is perhaps the intention of Blyth's quote, then Table 10 seems to be consistent with his claim, but the table has not been endorsed by his three conditions as a case of SP. Table 10 shows that even though there is no association between "gender" and "acceptance rates" of students in the overall school (column 8 in Table 10), there remains an association between "gender" and the "acceptance rates" when we divide the student population into two departments. In fact, we observe a clear interaction effect between "gender" and "departments" when the population has been partitioned into two departments.

Based on the information from the table together with his three conditions, we find that there is a tension between Blyth's conditions for SP and his understanding of SP in terms of interaction effects between variables. If we accept his three conditions for

SP then the example of showing the cessation of association between "gender" and "acceptance rates" gets eliminated as a possible example of SP. However, if we accept his interaction explanation for SP then the cessation of association has turned out to be compatible with his analysis of SP. This tension, however, does not arise for our account as the cessation of association between two variables shown by Table 10 can be subsumed as a case of SP.

So, primarily there are three fundamental differences between Blyth's account and ours. First, Blyth does not endorse the cessation of an association in the overall population as a case of SP. Second, we have already noted that there is a tension between consequences of his account and his preference to construe SP in terms of interactions effects of variables. Our third and final comment has to do with two key theorems of SP we have proved. They are (i) SP arises only if $A_1 \neq B_1$ and (ii) SP arises only if $A_2 \neq B_2$. Since they are theorems of SP they must hold for his version of SP. However, we are the ones who have pointed out these two theorems of SP. Based on these considerations, we reject the charge that our account of SP is no different from Blyth's. In fact, we have virtually argued that both are different accounts of SP.

7 Larger significance of our research

We began our paper by distinguishing among three types of questions. (i) Why or in what sense is Simpson's paradox a paradox? (ii) What is the correct analysis of the paradox? (iii) How one should proceed when confronted with a typical case of the paradox? Although these questions are no doubt distinct, our formal reconstruction of the paradox provides a unified account of them, which the empirical studies we have carried out illustrates and amplifies. We showed that Simpson's paradox can be generated in a straightforward deductive way. Among its premises, there is concealed a distinctly human dimension. In recent years, there has been a great deal of discussion of human frailty in connection with individuals' assessment of probabilistic statements. Our resolution of the paradox has illuminated another aspect of human frailty. We explained its apparent paradoxical nature by invoking the failure of our widespread intuitions about numerical inference. The failure of collapsibility in Simpson's paradox-type cases is what makes them puzzling, and the latter is what paints a human face onto the rather abstract structure of "Simpson's paradox."

Below we discuss the results of two experiments based on Simpson's paradox. One involves a version of the paradox in non-mathematical language and the second one is in mathematical language. The purpose of these experiments was to determine student responses to the following questions.

First experiment involves a non-mathematically explained case of the paradox:

Consider the following information to be correct.

There are only two high schools in a certain school district. Given that the graduation rate for girls in School #1 is higher than the graduation rate for boys in School 1, and that the graduation rate for girls in School #2 is higher than the graduation rate for boys in School 2. Does it follow that the graduation rate for girls in the district is higher than the graduation rate for boys in the district?

Which one of the following is true?

a. Yes, the graduation rate for girls is **greater than** it is for boys in the district.
b. No, the graduation rate for girls is **less than** it is for boys in the district.
c. No; the graduation rates for girls and boys are **equal** in the district
d. No inference could be made about the truth or falsity of the above because there is not enough information.

Second experiment involves a mathematically described case of the paradox:
Consider the following mathematical expressions to be correct.

1. $(f_1/F_1) > (m_1/M_1)$.
2. $(f_2/F_2) > (m_2/M_2)$.

Does it follow that $((f_1 + f_2)/(F_1 + F_2)) > ((m_1 + m_2)/(M_1 + M_2))$?
Which one of the following is true?

(a) Yes, the first expression is **greater than** the second.
(b) No, the first expression is **less than** the second.
(c) No, the first and second expressions are **equal**.
(d) No, inference could be made about the truth or falsity of the above because there is not enough information.

In an experiment with 106 student responses to both questions, we found that for the first non-mathematical question, students chose response (a) 83% of the time which involves the mistaken use of the collapsibility principle. They correctly responded choosing (d) only 12% of the time. For the mathematical question, they are right at the rate of 29%, whereas they have committed the error at 57% of the time.[14] Similar surveys over many years of students in philosophy classes have manifested the same patterns of responses.[15]

The math version of the paradox exactly mirrors our test case which does not involve any causal intuition whatsoever. In turn, the former also has the same structure as the non-math version of our experiment on the paradox. Consequently, it will be a mistake to think that the subjects' responses have exploited a causal intuition underlying different versions of the paradox based on the reason that there is no difference between these two experiments as they exhibit the same mathematical structure. Most subjects, as we have noticed erroneously took the inference in the Simpson's paradox type experiment to be inductive since they mistakenly applied the collapsibility principle (CP). In Simpson's paradox-type situations, subjects were confronted with possible choices in which those choices are subject to real-time constraints. The CP allows them to reach conclusions on the basis of the data quickly. That is to say, if subjects were to look at the data in the light of all possible information, they would perhaps never have reached a conclusion. In a broader perspective, we human beings are like

[14] Students chose the "correct" response in the formula-driven version of the question at a higher rate than in the story version of the question. This may be due to students' lack of certainty when presented with formulas than their ability to detect SP when provided with equations.

[15] This error is not unique to philosophy students. Two of our super-string theorist friends committed the same error when given these two experiments.

those subjects who are confronted with choices in their evolutionary history as well as their day to day life. If there is a trade-off between speed and error, reaching quick conclusions (on which our survival depends) will sometimes lead to error.

We have noticed that those experimental results show clearly that when confronted with a Simpson's paradox-type situation almost all subjects have jumped to an erroneous conclusion even though one of the options given to them in our experiment is to *not* make an inference. Why is this error committed so consistently when subjects are clearly given a choice not to make any inference in that situation? One plausible suggestion is that confronted with Simpson's paradox-type situations, the pressing issue for subjects is to make a decision rather than to suspend inference. On this suggestion, the "what to do" question seems more pressing under uncertainty in at least many situations. According to our analysis, however, this error is the misuse of the CP across the board and has nothing to do with the "what to do question". The result of our analysis is to divorce the question of the paradox and the reason it seems paradoxical from the question of the solution to the "what to do question." Most causal theorists including Pearl think that the latter has a resolution in some sort of causal analysis. Our account does not say that this is not possible. But a causal analysis of the "what to do question" should relate to cases where correlations are confused with causations, whereas the discussion of the paradoxical nature of Simpson's data sets should be related to other mathematical mistakes that people are prone to make that lead them into trouble.

It would be a mistake to assume that scholars in general completely agree with our account of the paradox. Steven Sloman, a cognitive scientist, wrote (in an email communication, April 24, 2009) "I [Sloman] believe that this paper is addressing a fundamental psychological question that I generally frame in terms of outside vs. inside perspectives (closely related to extensional vs. intentional perspectives). Contra Pearl and I, you are arguing that people's reasoning is from the outside, in terms of proportions, whereas we argue that it's from the inside, in terms of causal structure. . . . I note that these are not mutually exclusive perspectives and that each could capture different aspects of human reasoning."[16] There is no reason to disagree that people's reasoning is often from the 'inside.' Our point is simply that the reasoning does not have to be from the 'inside' for the aura of paradox to be generated. In our day to day situations, we humans look for deeper 'causal structures,' and are puzzled by our inability, as in Simpson-type cases, to find them. But we are more often puzzled, or so we have argued here, by the fact that such deeply held inductive habits as the principle of collapsibility lead to paradoxical conclusions. The authors are Humeans to this extent that we prefer explanations of untoward results in terms of habits rather than of 'causality.' The latter tend to substitute one mystery for another. We believe, however, that better to stay on the 'outside' where the sun shines.

Acknowledgments We would like to thank John G. Bennett, John Borkowski, Robert Boik, Martin Curd, Dan Flory, Debzani Deb, David Freedman, Caleb Galloway, Jack Gilchrist, Sander Greenland, Martin Hamilton, Joseph Hanna, Daniel Hausman, Christopher Hitchcock, Jayanta Ghosh, Autumn Laughbaum, Adam Morton, Dibyendu Nandi, Michael Neeley, Daniel Osherson, James Robison-Cox, Prasun Roy, Federica

[16] For Sloman's view on the distinction between the inside and outside perspectives, see Sloman (2005).

Russo,Tasneem Sattar, Billy Smith, Elliott Sober, Steve Sloman, Steve Swinford, Olga Vsevolozhskaya, and Paul Weirich for comments on an earlier version of the paper. We also thank Paul Humphreys for calling our attention to some references relevant to the paper and Donald Gillies for his encouragement to consider the paradox the way we have done it in this paper. The paper also benefited from the comments received at two workshops at the University of Konstanz, a conference at the University of Alabama in Birmingham, Montana Chapter of the *American Statistical Association* meetings in Butte, *Society for the Exact Philosophy meetings* at the University of Alberta, *Indian Institute of Science and Educational Research* in Mohanpur, *Centre for Philosophy and Foundations of Science* in New Delhi, Statistics Colloquium at Montana State University, Bhairab Ganguly College in Kolkata, an international conference on "Scientific Methodology" at Visha-Bharati University and Ecology Seminar at Montana State University where the paper was presented. We are also thankful to several referees of various journals including the referees of this journal for their useful comments regarding the paper. Special thanks are to John G. Bennett for his numerous insightful suggestions regarding the content of the paper. The research for this paper has been supported both by the *NASA's* Astrobiology Research Center (grant #4w1781) and the Research and Creativity grant from Montana State University.

Appendix

For proving Theorems 1 to 3 and lemmas 1 to 4, we have used two assumptions (call them a and b, respectively and two definitions ("α" and "β"). We have defined α and β differently than what we have done previously in the text.

1. Let "a" = (members of A partition in 1)/(total members of A)
2. Let "b" = (members of B partition in 1)/(total members of B)

To give an intuitive feeling of what a and b stand for, we could use Simpson's paradox type I in which a = 200/500 and b = 600/700.

Let the quantities A_1, A_2, B_1, B_2, a, b be in [0,1], where A_1, A_2, B_1, and B_2 are as before, where for example, A_1 is the ratio of the number of females accepted in Department 1 to the total number of females that applied to Department 1 and B_2 is the ratio of the number of males accepted in Department 2 to the total number of males that applied to Department 2.

Define $\alpha := aA_1 + (1 - a)A_2$ and $\beta = bB_1 + (1 - b)B_2$. The Simpson Paradox results from two conditions being imposed on the the above quantities:

(i) Condition 1 (C_1) : $A_1 \geq B_1$;
(ii) Condition 2 (C_2) : $A_2 \geq B_2$;
(iii) Condition 3 (C_3) : $\beta \geq \alpha$;
(iv) Condition 4 (C_4) : $\theta = (A_1 - B_1) + (A_2 - B_2) + (\alpha - \beta) > 0$.

Theorem 1 *Simpson's Paradox results only if* $A_1 \neq A_2$.

Proof Assume to the contrary that $A_1 = A_2$. We force a contradiction of condition (C_3). First, given our assumption,

$$\alpha = aA_1 + (1 - a)A_2 = aA_1 + (1 - a)A_1 = aA_1 + A_1 - aA_1 = A_1 = A_2.$$

There are three cases to consider with respect to B_1 and B_2.

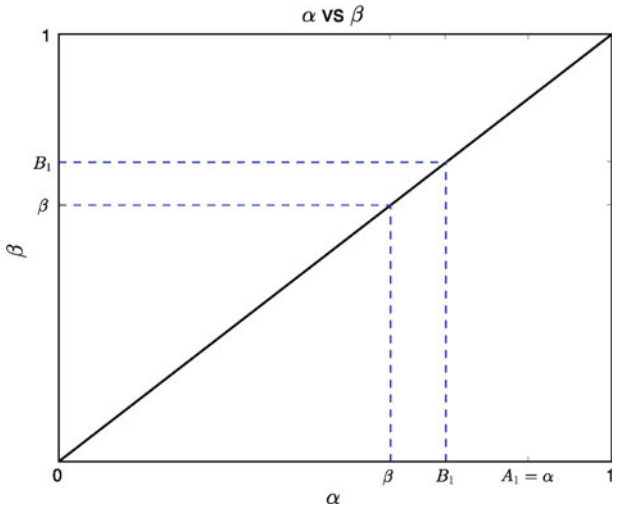

Fig. 1 Graph illustrating proof of (i)

(i) Suppose that $B_1 > B_2$. In this case, the following relationships hold amongst B_1, B_2, and β:

$$bB_1 + (1 - b)\, B_1 > bB_1 + (1 - b)\, B_2 = \beta \Rightarrow B_1 > \beta.$$

Geometrically, this places B_1 above β. See Fig. 1, where the vertical axis represents β and the horizontal axis represents α. We use the line $\alpha = \beta$ to place B_1 and β on the α-axis in order to compare them to A_1 and α. By (C_1) $A_1 \geq B_1$, which forces the inequalities $\beta < B_1 \leq A_1 = \alpha$. This contradicts condition (C_3). The relative positions of these quantities are illustrated on the horizontal axis of Fig. 1.

(ii) Suppose $B_2 > B_1$. As in (i), $bB_2 + (1 - b)B_2 > bB_1 + (1 - b)B_2 = \beta \Rightarrow B_2 > \beta$. Using the relation $A_2 \geq B_2$ from (C_2), we conclude similarly that $\beta < B_2 \leq A_2 = \alpha$, contradicting (C_3). (To see this in Fig. 1, replace B_1 by B_2 and A_1 by A_2.)

(iii) Lastly, take $B_1 = B_2$. As above, $A_1 = \alpha$ and $B_1 = \beta$. From the assumption that $A_1 \geq B_1$ we have that $\alpha \geq \beta$. And by (C_3) we have that $\alpha \leq \beta$. The only way for both inequalities to hold is for $\alpha = \beta$. Yet if this is the case then $\theta = 0$, contradicting (C_4). Therefore, equality of A_1 and A_2 is incompatible with Simpson's Paradox.

Theorem 2 *Simpson's paradox arises only if* $B_1 \neq B_2$.

We first prove a lemma.

Lemma 1: The following relationships hold:

LM1: If $A_1 > A_2$, *then* $A_1 > \alpha > A_2$.
LM2: If $A_2 > A_1$, *then* $A_2 > \alpha > A_1$.
LM3: If $B_1 > B_2$, *then* $B_1 > \beta > B_2$.
LM4: If $B_2 > B_1$, *then* $B_2 > \beta > B_1$.

Proof These relationships are symmetric with respect to the indices. We prove only (1) and (3). The other cases are handled by swapping variables and indices.

LM1: We have made use of the following algebraic identity already: $A_1 = aA_1 + A_1 - aA_1 = aA_1 + (1 - a)A_1$. For case (1), where $A_1 > A_2$ it follows that

$$A_1 = aA_1 + (1 - a)A_1 > \alpha = aA_1 + (1 - a)A_2 > aA_2 + (1 - a)A_2 = A_2.$$

Hence, if $A_1 > A_2$, then $A_1 > \alpha > A_2$.

LM3: Similarly, when $B_1 > B_2$ as in case (ii),

$$B_1 = bB_1 + (1 - b)B_1 > \beta = bB_1 + (1 - b)B_2 > bB_2 + (1 - b)B_2 = B_2.$$

Thus, whenever $B_1 > B_2$, we can bound β by $B_1 > \beta > B_2$.

Proof of Theorem 2 We proceed by supposing that $B_1 = B_2$. From the algebraic identity used in the above lemma it follows that $\beta = B_1 = B_2$. Since we have shown that A_1 cannot equal A_2, we assume without loss of generality that $A_1 > A_2$; thus $A_1 > \alpha > A_2$ by the lemma. Yet, by condition C_2, $A_2 \geq B_2$. This implies that $\alpha > A_2 \geq B_2 = \beta$. In particular this forces $\alpha > \beta$, contradicting the reversal of C_3. Thus, the case where $B_1 = B_2$ and $A_1 > A_2$ cannot arise. In Fig. 2, the case where A_2 is strictly greater than B_2 is shown.

If instead $A_2 > A_1$, then we switch the A's to conclude that $\alpha > A_1 \geq \beta = B_2$. Therefore, it is incompatible with Simpson's Paradox for B_1 to equal B_2.

Theorem 3 *Simpson's paradox arises only if* $(A_1 \neq A_2)$ *if and only if* $(B_1 \neq B_2)$.

By definition, Theorem 3 says Simpson's paradox arises only if $\{(A_1 \neq A_2) \Rightarrow (B_1 \neq B_2)\}$ *and* $\{(B_1 \neq B_2) \Rightarrow (A_1 \neq A_2)\}$.

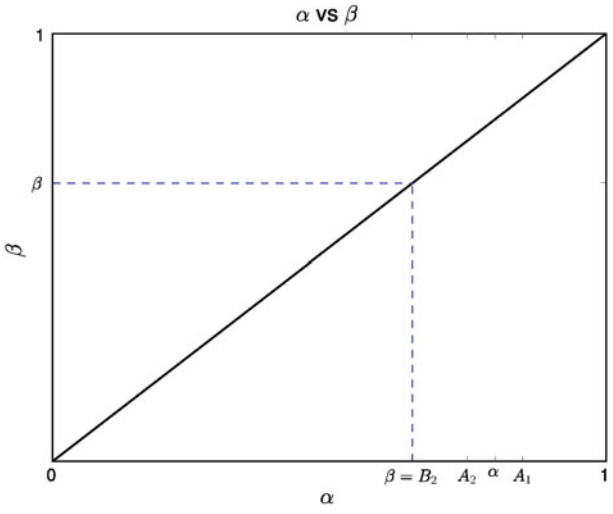

Fig. 2 Graph showing the case when $B_2 = \beta < A_2 < \alpha < A_1$

Proof Consider the first conjunct $(A_1 \neq A_2) \Rightarrow (B_1 \neq B_2)$. This condition is logically equivalent to $(B_1 = B_2) \Rightarrow (A_1 = A_2)$. The antecedent of that conditional is false because of Theorem 2. Therefore, $(B_1 = B_2) \Rightarrow (A_1 = A_2)$ is true, which proves $(A_1 \neq A_2) \Rightarrow (B_1 \neq B_2)$. The proof for $(B_1 \neq B_2) \Rightarrow (A_1 \neq A_2)$ is similar to the first conjunct. $\{(B_1 \neq B_2) \Rightarrow (A_1 \neq A_2)\}$ is logically equivalent to $\{(A_1 = A_2) \Rightarrow (B_1 = B_2)\}$. The latter is true because the antecedent of the conditional is false by Theorem 1. Therefore, the conjunction is true, leading to Theorem 3.

Bibliography

Blyth, C. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366 (Theory and Method Section).

Cartwright, N. (1979). Causal laws and effective strategies. *Nous, 13*, 419–437.

Cartwright, N. (1999). *The dappled word: A study of the boundaries of science*. UK, Cambridge: Cambridge University Press.

Clark, M. (2002). *Paradoxes from A to Z*. London: Routledge.

Elles, E., & Sober, E. (1983). Probabilistic causality and the question of transitivity. *Philosophy of Science, 50*, 35–57.

Freedman, D., Pisani, R., & Purve, R. (1999). *Statistics* (3rd ed.). New York: W. W. Norton & Company.

Good, I. J., & Mittal, Y. (1988). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics, 15*(2), 694–711.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 19*, 29–46.

Hausman, D. (1998). *Causal asymmetries*. Cambridge: Cambridge University Press.

Hoover, K. (2001). *Causality in microeconomics*. England: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and basics*. England: Cambridge University Press.

Kyburg, H. (1997). The rule of adjunction and reasonable inference. *Journal of Philosophy, XCIV*(3), 109–125.

Lindley, D., & Novick, M. (1981). The role of exchangeability in inference. *Annals of Statistics, 9*(1), 45–58.

Malinas, G. (2001). Simpson's paradox: A logically benign, empirically treacherous hydra. *The Monist, 84*(2), 265–283.

Meek, C., & Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science, 45*, 1001–1021.

Mittal, Y. (1991). Homogeneity of subpopulations and Simpson's paradox. *Journal of the American Statistical Association, 86*, 167–172.

Morton, A. (2002). If you're so smart why are you ignorant? Epistemic causal paradoxes. *Analysis, 62*(2), 110–116.

Novick, M. R. (1983). The centrality of Lord's paradox and exchangeability for all statistical inference. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement*. Hillsdale, NJ: Erlbaum.

Otte, R. (1985). Probabilistic causality and Simpson's paradox. *Philosophy of Science, 52*(1), 110–125.

Pearl, J. (2000). *Causality*. (1st ed.). Cambridge: Cambridge University Press

Pearl, J. (2009). *Causality*. (2nd ed.). Cambridge: Cambridge University Press

Rothman, K., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott Williams.

Savage, L. (1954). *Foundations of statistics*. New York: Wiley.

Simpson, H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series. B, 13*(2), 238–241.

Skyrms, B. (1980). *Causal necessity*. New York: Yale University Press.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Sober, E., & Wilson, D. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Mass: Harvard University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge: MIT Press.