Simpson's Paradox: A Singularity of Statistical and Inductive Inference

Palash Sarkar
Applied Statistics Unit
Indian Statistical Institute
203, B.T. Road, Kolkata
India 700108
email: palash@isical.ac.in

Prasanta S. Bandyopadhyay
Department of History & Philosophy
Montana State University
Bozeman, MT
USA
email: psb@montana.edu

April 1, 2021

Abstract

The occurrence of Simpson's paradox (SP) in 2×2 contingency tables has been well studied in the literature. The first contribution of the present work is to comprehensively revisit this problem. We provide a threadbare analysis of SP in 2×2 contingency tables and present new results, detailed proofs of previous results and a unifying view of the important examples of SP that have been reported in the literature. The second contribution of the paper suggests a new perspective on the surprise element of SP, raises some critical questions for the causal analysis of SP and provides a broad perspective on logic, probability and statistics with SP at its centre.
Keywords: Simpson's paradox, contingency tables, odds ratio, causal graphs.

1 Introduction

Statistical inference analyses data and provides information about the underlying population. In many settings, the population may be considered to consist of two or more separate sub-populations. Data corresponding to individual sub-populations would be separately available and the data for the overall population may be obtained by aggregating the sub-population data. Such aggregation of sub-population data can have consequences that are hard to interpret in statistical terms. In particular, the possibility of cessation or reversal of association/preference/relation existing in each of the sub-population data upon aggregation appears to be paradoxical. Such observations were reported more than a century ago by Pearson et al. (1899) and Yule (1903). Presently, such a property is called Simpson's paradox (SP) since Blyth (1972) so named it based on a paper by Simpson (1951). It has also been called the Yule reversal paradox in Mittal (1991). For a list of early works where similar reversals have been reported, we refer to the introduction of the paper by Armistead (2014). SP has continued to fascinate a long line of researchers. In particular, attempts have been made to explain the paradox in terms of other concepts such as homogeneity (Mittal (1991)), causality (Pearl (2009, 2014)), and confirmation theory (Fitelson (2017)). Another direction of research has uncovered Simpson's paradox in diverse areas such as quantum mechanics and measurement of economic inequality (see Selvitella (2017, 2017a) and Chakravarty and Sarkar (2020)).

Statistical, and more generally inductive, inference is arguably one of the most important methods of knowledge acquisition. SP with its uniquely distinctive characteristic remains a sigularity in the theory of statistical and inductive inference. Our work continues the long line of research in elucidating and understanding SP from both mathematical and philosophical points of view.

The first contribution of the paper is to take a fresh look at SP arising in a pair of 2×2 contingency tables. The overview of SP in the Stanford Encyclopedia of Philosophy by Sprenger and Weinberger (2021) also consider SP in terms of 2×2 contingency tables and briefly discuss some of the results that have appeared in the literature. We make a deeper analysis of SP in the setting of 2×2 contingency tables.

The mathematical machinery that we use is based on probability and logic. The literature provides several slightly different definitions of SP. We make a comprehensive analysis of all the possible cases that can arise from two 2×2 contingency tables and based on this we arrive at a complete formal definition of SP. The subsequent task is to analyse the definition so as to obtain conditions for SP to hold. In this exercise, it turns out that the Odds Ratio plays an important role. We provide a simple characterisation of SP in terms of the Odds Ratio. Following up on the work of Good and Mittal (1987), we introduce the notion of weak Odds Ratio homogeneity and show that this constitutes a sufficient condition for SP not to hold. A homogeneity condition was introduced by Mittal (1991) and shown to be sufficient for SP not to hold. We provide a new and simpler proof of this result based on the Odds Ratio characterisation of SP. The important notion of positive association of random variables was introduced by Lindley and Novick (1981). We provide a characterisation of this notion in terms of Odds Ratio. Using the language of positive association, Lindley and Novick (1981) had stated a necessary condition for SP. This result was expanded by Mittal (1991). We provide a new proof of the expanded result which fills in some important details that are missing from Mittal's proof.

Statistical inference is an inductive procedure. With progressive accumulation of data, it is possible that SP can arise and vanish over time. We prove that this can indeed happen where the appearance and disappearance of SP can toggle ad infinitum. This result may be interpreted in the broader context of the recent account of inductive logic which is based on conditional probabilities to quantify the degree of support that evidence statements provide to hypotheses (see Hawthorne (2021)).

We provide a unifying description of some of the well known examples of SP that have appeared in the literature. In describing these examples, we distinguish the numerical or the data oriented part from the contextual or the interpretative part. This leads to the intriguing possibility that the contextual part of one example can be combined with the numerical part of another example to create a new example of SP. An explanation or resolution of SP in one combination of numerical-plus-contextual example may not hold in another such combination. This shows that any explanation of SP based only on the contextual part is likely to be unsatisfactory.

Any paradox has a surprise element. From a philosophical point of view, it is of interest to understand the reason behind the surprise. The second contribution of the paper is to provide a different perspective on the surprise element of SP. We argue that SP can be viewed in the general framework of paradoxes where human intuition expects the whole to behave in the same manner as the parts. The credibility of such a viewpoint arises from the existence of the so-called aggregation paradoxes. Pearl (2014) had suggested that the only way to view the surprise element of SP is in terms of causality. We point out two objections to this line of thought.

Pearl's work on causal analysis of SP (Pearl (2009, 2014)) has been influential and is widely referred in the literature. We raise some issues regarding this analysis. In particular, we point out that there are examples of SP which do not have a natural causal interpretation. So, causal modelling cannot cover all examples of SP. More generally, we highlight that statistical inference is an inductive inference procedure while the framework of causal structure is essentially a deductive theory. Conflation of the two branches lead to some well known foundational issues.

A key question when faced with an instance of SP, is whether to accept the conclusion of the sub-population data or to accept the conclusion of the aggregated data. It has been suggested by Pearl (2014) that depending on the situation, either the conclusion of the sub-population data or the conclusion of the aggregated data, or, neither should be accepted. We make two points regarding this suggestion. First, that that there can be situations where both the conclusions, i.e., the conclusions of both the aggregated and the disaggregated tables, can be meaningful. Second, that among the examples of SP that appear in the literature, there is none where the conclusion of both the aggregated and the disaggregated data are to be rejected, raising the question of whether the possibility of rejecting both conclusions is only a mathematical artifice. While Pearl had suggested causal structures to be the definitive way of resolving SP, in contrast to his suggestion, our final conclusion is that rather than considering causal structure modelling to be the definitive tool, it should properly be considered as one of the tools in the toolkit of a statistician who is faced with an instance of SP and wishes to take a decision.

In the final section, we provide a broad discussion of the interplay of logic, probability and philosophy revolving around causality. The discussion is anchored in philosophy of statistics with SP at its core. We put forward the view that an appropriate amalgamation of statistical and causal techniques can lead to improved methods of scientific inference.

2 Definition of Simpson's Paradox

In this section, we take a detailed look at what constitutes a definition of Simpson's Paradox in terms of 2×2 contingency tables.

Let T be a 2×2 contingency table of the form given below.

$$T = \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array}. \tag{1}$$

We will assume that a, b, c and d are positive real numbers.

Given a contingency table T as in (1), one may associate true/false valued random variables X, Y and $\overline{X}$ and $\overline{Y}$ as follows: X and $\overline{X}$ correspond to the first and second row respectively of T; Y and $\overline{Y}$ correspond to the first and second column respectively of T. Note that X (resp. Y) is true if and only if $\overline{X}$ (resp. $\overline{Y}$) is false. So, $\Pr[Y|X] = a/(a+b)$ and $\Pr[Y|\overline{X}] = c/(c+d)$ and similarly for the other conditional probabilities. Here the notation $\Pr[Y|X]$ is implicitly used to denote the probability of the event $\Pr[Y = \text{true}|X = \text{true}]$ and similarly for other probabilities.

Definition 1. *Let T_1 and T_2 be two 2×2 contingency tables as shown below.*

$$T_1 = \begin{array}{|c|c|} \hline a_1 & b_1 \\ \hline c_1 & d_1 \\ \hline \end{array}, \quad T_2 = \begin{array}{|c|c|} \hline a_2 & b_2 \\ \hline c_2 & d_2 \\ \hline \end{array}.$$

The collapsed (or aggregated) table $T_1 + T_2$ is the following.

$$T_1 + T_2 = \begin{array}{|c|c|} \hline a_1 + a_2 & b_1 + b_2 \\ \hline c_1 + c_2 & d_1 + d_2 \\ \hline \end{array}.$$

Notation: We fix the following notation for later use.

$$\alpha_1 = a_1 + b_1; \quad \alpha_2 = a_2 + b_2; \quad \gamma_1 = c_1 + d_1; \quad \gamma_2 = c_2 + d_2;$$
$$A_1 = \frac{a_1}{\alpha_1}; \quad A_2 = \frac{a_2}{\alpha_2}; \quad C_1 = \frac{c_1}{\gamma_1}; \quad C_2 = \frac{c_2}{\gamma_2};$$
$$\mu = \frac{a_1 + a_2}{\alpha_1 + \alpha_2}; \quad \nu = \frac{c_1 + c_2}{\gamma_1 + \gamma_2}.$$

Figure 1: Classifications of the 27 relationships that can occur for the tables T_1, T_2 and $T_1 + T_2$.

Case	cond on T_1	cond on T_2	cond on $T_1 + T_2$	remark
1	$A_1 > C_1$	$A_2 > C_2$	$\mu > \nu$	Aligned
2	$A_1 > C_1$	$A_2 > C_2$	$\mu = \nu$	Weak paradox
3	$A_1 > C_1$	$A_2 > C_2$	$\mu < \nu$	Paradox
4	$A_1 > C_1$	$A_2 = C_2$	$\mu > \nu$	Class-2
5	$A_1 > C_1$	$A_2 = C_2$	$\mu = \nu$	Class-3
6	$A_1 > C_1$	$A_2 = C_2$	$\mu < \nu$	Class-4
7	$A_1 > C_1$	$A_2 < C_2$	$\mu > \nu$	Class-1
8	$A_1 > C_1$	$A_2 < C_2$	$\mu = \nu$	Class-5
9	$A_1 > C_1$	$A_2 < C_2$	$\mu < \nu$	Class-1
10	$A_1 = C_1$	$A_2 > C_2$	$\mu > \nu$	Class-2
11	$A_1 = C_1$	$A_2 > C_2$	$\mu = \nu$	Class-3
12	$A_1 = C_1$	$A_2 > C_2$	$\mu < \nu$	Class-4
13	$A_1 = C_1$	$A_2 = C_2$	$\mu > \nu$	Class-6
14	$A_1 = C_1$	$A_2 = C_2$	$\mu = \nu$	Class-0
15	$A_1 = C_1$	$A_2 = C_2$	$\mu < \nu$	Class-6
16	$A_1 = C_1$	$A_2 < C_2$	$\mu > \nu$	Class-4
17	$A_1 = C_1$	$A_2 < C_2$	$\mu = \nu$	Class-3
18	$A_1 = C_1$	$A_2 < C_2$	$\mu < \nu$	Class-2
19	$A_1 < C_1$	$A_2 > C_2$	$\mu > \nu$	Class-1
20	$A_1 < C_1$	$A_2 > C_2$	$\mu = \nu$	Class-5
21	$A_1 < C_1$	$A_2 > C_2$	$\mu < \nu$	Class-1
22	$A_1 < C_1$	$A_2 = C_2$	$\mu > \nu$	Class-4
23	$A_1 < C_1$	$A_2 = C_2$	$\mu = \nu$	Class-3
24	$A_1 < C_1$	$A_2 = C_2$	$\mu < \nu$	Class-2
25	$A_1 < C_1$	$A_2 < C_2$	$\mu > \nu$	Paradox
26	$A_1 < C_1$	$A_2 < C_2$	$\mu = \nu$	Weak paradox
27	$A_1 < C_1$	$A_2 < C_2$	$\mu < \nu$	Aligned

We connect the above quantities to conditional probabilities. Let M and $\overline{M}$ be true/false valued random variables associated with tables T_1 and T_2 respectively; X and $\overline{X}$ (resp. Y and $\overline{Y}$) correspond to the first and second rows (resp. columns) of T_1, T_2 and $T_1 + T_2$. Then $\Pr[Y|X, M] = A_1$, $\Pr[Y|\overline{X}, M] = C_1$, $\Pr[Y|X, \overline{M}] = A_2$, $\Pr[Y|\overline{X}, \overline{M}] = C_2$, $\Pr[Y|X] = \mu$ and $\Pr[Y|\overline{X}] = \nu$. Note that the individual tables T_1 and T_2 provide the joint probabilities $X \wedge Y$, $X \wedge \overline{Y}$, $\overline{X} \wedge Y$, $\overline{X} \wedge \overline{Y}$ conditioned on M and $\overline{M}$ respectively, whereas the merged table $T_1 + T_2$ provides the unconditional probabilities of these four events.

From the tables T_1, T_2 and $T_1 + T_2$, there are 27 relationships that arise. These are summarised in Figure 1. The last column of the table provides a classification of these relationships. The basis for this classification is as follows.

Class-0: Equality holds in all the three tables T_1, T_2 and $T_1 + T_2$.

Aligned: The inequalities in the Tables T_1, T_2 and $T_1 + T_2$ are strict and are all in the same direction.

Paradox: The inequalities in the Tables T_1, T_2 and $T_1 + T_2$ are strict; the inequalities in the Tables T_1 and T_2 are in the same direction, while the inequality in the Table $T_1 + T_2$ is in the other direction.

Weak paradox: The inequalities in the Tables T_1 and T_2 are strict and are in the same direction, while there is equality in Table $T_1 + T_2$.

Class-1: The inequalities in the Tables T_1, T_2 and $T_1 + T_2$ are strict. The inequalities in T_1 and T_2 are in opposite directions.

Class-2: One of the inequalities in Tables T_1 and T_2 is strict, while the other is a equality. The inequality in Table $T_1 + T_2$ is strict and is in the same direction as the strict inequality in T_1 or T_2.

Class-3: One of the inequalities in Tables T_1 and T_2 is strict, while the other is a equality. Equality holds in Table $T_1 + T_2$.

Class-4: One of the inequalities in Tables T_1 and T_2 is strict, while the other is a equality. The inequality in Table $T_1 + T_2$ is strict and is in the opposite direction to the strict inequality in T_1 or T_2.

Class-5: The inequalities in the Tables T_1 and T_2 are strict and in the opposite directions. Equality holds in Table $T_1 + T_2$.

Class-6: Equality holds in the Tables T_1 and T_2, while strict inequality holds in Table $T_1 + T_2$.

A relevant question is whether some of the cases in Figure 1 are logically impossible, i.e., whether there is a case for which the conditions are such that no pair of 2×2 contingency tables satisfy the conditions? We show that the answer to this question is negative.

Theorem 1. *All the cases considered in Figure 1 are logically possible.*

Equality represents a no-preference condition, while a strict inequality represents a clear preference of one row over the other. Let us try to understand the entries in Figure 1 in terms of the interpretation of equality as no-preference and strict inequality as a clear preference. The understanding of Class-0 and the 'Aligned' class are clear. In Class-0, all the three tables represent no-preference conditions, while in the 'Aligned' class, the preferences in the individual tables are aligned with that in the aggregated table. If the condition of Class-0 occurs, then it can be said with strong confidence that the data does not provide any choice. If the condition of 'Aligned' class occurs, then a choice can be made without any ambiguity. Consider the cases of 'Weak paradox'. The inequalities in the Tables T_1 and T_2 are strict and in the same direction, but equality holds in the Table $T_1 + T_2$. So, it can be said that the preference which can be seen in the individual tables disappears in the aggregated table. The more vexing scenario is the cases of 'Paradox'. The inequalities in the two individual tables are strict and in the same direction, while the strict inequality in the aggregated table is in the opposite direction. This creates the paradox of reversal, where a choice which is individually better in the sub-population becomes worse-off in the aggregated population. Making a proper choice in the face of this paradox is a difficult problem.

Informally, Simpson's paradox can be understood to be the following statement covering the classes 'Weak paradox' and 'Paradox'.

> Both the sub-populations provide a clear preference to one particular choice, but this preference is either reversed or ceases to exist in the entire population.

Let us now consider the formal definitions of Simpson's paradox in Blyth (1972), Mittal (1991) and Bandyopadhyay et al. (2011).

B72: $(A_1 \geq C_1) \wedge (A_2 \geq C_2) \wedge (\mu < \nu)$ (Blyth (1972)).

M91: $((A_1 \geq C_1) \wedge (A_2 \geq C_2) \wedge (\mu \leq \nu)) \vee ((A_1 \leq C_1) \wedge (A_2 \leq C_2) \wedge (\mu \geq \nu))$ (Mittal (1991).

BNGBB11: $(A_1 \geq C_1) \wedge (A_2 \geq C_2) \wedge (\nu \geq \mu)$ and at least one of the inequalities is strict. (Bandyopadhyay (2011)).

The condition M91 is given in Mittal (1991) in a different, but equivalent form. The condition B72 captures the Cases 3, 6, 12, and 15, the condition M91 captures the Cases 2, 3, 5, 6, 11 to 17, 22, 23, 25, 26, and the condition BNGBB11 captures the Cases 2, 3, 5, 6, 11, 12, and 15.

The condition M91 covers both the classes 'Weak Paradox' and 'Paradox'. On the other hand, both B72 and BNGBB11 are incomplete, since neither of them capture Case 25 which is a clear case of the paradox. Further, B72 also does not cover Cases 2 and 26, where a clear preference in the sub-populations ceases to exist in the aggregated population. The conditions B72 and BNGBB11 can be expanded. Consider the following conditions.

B72$'$: $(A_1 \leq C_1) \wedge (A_2 \leq C_2) \wedge (\mu > \nu)$.

BNGBB11$'$: $(A_1 \leq C_1) \wedge (A_2 \leq C_2) \wedge (\nu \leq \mu)$ and at least one of the inequalities is strict.

Then Exp-B72 : B72$\vee$B72$'$ is the expanded form of Blyth's condition and Exp-BNGBB11 : BNGBB11$\vee$ BNGBB$'$ is the expanded form of Bandyopadhyay et al.'s condition. We note that Exp-B72 still does not cover Cases 2 and 26 and hence is still incomplete.

In B72, Exp-B72, BNGBB11, Exp-BNGBB11 and M91, the conditions on the sub-populations (i.e., in tables T_1 and T_2) include the equality condition. As a consequence, certain cases get included in the definition which would not be part of the informal description of Simpson's paradox mentioned above. For example, Case 15, i.e., $A_1 = C_1$, $A_2 = C_2$ and $\mu < \nu$ is covered by both B72 and BNGBB11. This case states that both the sub-populations indicate no preference while a preference is indicated by the whole population. While this may be considered to be surprising, it is certainly not a case where both the sub-populations show a clear trend which is reversed or ceases to exist in the entire population.

The condition M91 has equality conditions on both the sub-populations as well as the aggregated population. This results in M91 covering the Case 14, where equality holds for both the sub-populations as well as the aggregate population. This clearly cannot be considered to be surprising.

Based on the above discussion, we formulate the following definition to exactly cover the informal statement of Simpson's paradox.

Definition 2. *Given two 2×2 contingency tables T_1 and T_2, Simpson's paradox (SP) is said to occur for T_1 and T_2 if* $\mathsf{SP} : \mathsf{SP}_1 \vee \mathsf{SP}_2$ *holds, where*

SP_1*:* $(A_1 > C_1) \wedge (A_2 > C_2) \wedge \neg(\mu > \nu)$,

SP_2*:* $(A_1 < C_1) \wedge (A_2 < C_2) \wedge \neg(\mu < \nu)$.

The condition SP_1 captures Cases 2 and 3 while SP_2 captures Cases 25 and 26 of Figure 1. Suppose SP_2 holds. Interchanging the first and second rows of the tables T_1, T_2 and T_3, we see that SP_1 holds. So, an interchange of the rows of the tables converts SP_2 to SP_1. Further, both SP_1 and SP_2 cannot simultaneously hold.

Since Exp-B72 does not cover the cases of 'Weak paradox', we make further comparison only with M91 and Exp-BNGBB11. Both our definition and Exp-BNGBB11 do not cover the classes labelled 'Aligned' and the Classes-0, 1, 2 and 5, i.e., cases in these classes are not considered to be examples of SP by either our definition or under Exp-BNGBB11. The difference between Exp-BNGBB11 and our definition is that Exp-BNGBB11 includes Classes 3, 4 and 6, while our definition does not. The difference between M91 and our definition is that M91 includes Classes 0, 3, 4 and 6, while our definition does not. As mentioned earlier, Class 0 is clearly not paradoxical. Next we contrast some of the classes which are covered by M91 and Exp-BNGBB11 with some of those which are not covered by these conditions.

1. Contrast Class-4 with Class-1. The similarity between these two classes is that the inequality in the aggregated population is strict and is in opposition to the strict inequality in one of the sub-populations; the dissimilarity is that in Class-1, the inequality in the other sub-population is strict and in reverse direction to that in the aggregated population, while in Class-4, equality holds in the other sub-population. This viewpoint suggests that if Class-1 is not viewed as a case of SP, then neither should Class-4 be.

2. In Class-3, there is a strict inequality in exactly one of the sub-populations, while equality holds in the other as well as the aggregated population. So, a clear preference indicated by exactly one of the sub-populations ceases to exist in the aggregated population. Again, we contrast with Class-1, where a clear preference indicated by one of the sub-populations is reversed in the aggregated population. So, if Class-1 is not considered to be a case of SP, neither should Class-3 be.

3. In Class-6, both the sub-populations indicate a no-preference condition, while the aggregated population shows a clear preference. Depending upon the context, this may or may not be surprising. On the other hand, Class-6 is clearly not covered by the informal statement describing SP.

From Theorem 1, it follows that all of the cases in Table 1 are logically possible. Though possible, not all of them are equally likely. In particular, the equality conditions in T_1, T_2 and $T_1 + T_2$ are perhaps less likely and hence quite unlikely to occur in practice. At this point though, we do not have a formal proof that the equality conditions for T_1, T_2 and $T_1 + T_2$ are less likely to occur compared to the other cases. If we indeed assume that equality conditions do not occur in any of these tables, then the definition of SP that we have formulated is equivalent to Exp-B72, M91 and Exp-BNGBB11.

We have considered SP from the viewpoint of aggregating two contingency tables. The concept easily extends to more than two tables. Our definition of SP has formalised the notion of cessation or reversal of relation upon aggregation. Interestingly, a scenario of Simpson's paradox has been considered to arise even without cessation or reversal. Rather, it is based on the magnitudes of the relevant probabilities. For example, if $A_1 \gg C_1$ and $A_2 \gg C_2$, and $\mu > \nu$, but not $\mu \gg \nu$, then the strength of the relation in the two sub-populations has been reduced upon aggregation. Such an effect has been considered to be paradoxical. See for example Page 376 of Pagano and Gauvreau (1993). A formal definition of SP which captures the magnitude effect can be developed and the consequent results analysed. We leave such work for the future.

3 Necessary and/or Sufficient Conditions for Simpson's Paradox

Theorem 2 below provides necessary conditions for SP to hold and consequently, the negations of these conditions constitute sufficient conditions for SP not to hold.

Theorem 2. *Suppose* SP *occurs for* T_1 *and* T_2*. Then the following conditions hold.*

1. $A_1 \neq A_2$.
2. $C_1 \neq C_2$.
3. $\alpha_1 \neq \gamma_1$ *or* $\alpha_2 \neq \gamma_2$.

We note that in TH1 and TH2, Bandyopadhyay et al. (2011) had proved that the Condition BNGBB11 implies $A_1 \neq A_2$ and $C_1 \neq C_2$. The proofs that we provide for the first two points of Theorem 2 are simpler than the proofs of TH1 and TH2 in Bandyopadhyay et al. (2011).

A concept that turns out to be useful in our study of SP is the well known Odds Ratio which is defined as follows..

Definition 3 (Odds Ratio)**.** *The Odds (or, Cross-Product) Ratio* $\kappa(T)$ *for a* 2×2 *contingency table as in* (1) *is defined to be* $\kappa(T) = (ad)/(bc)$.

The following result provides a characterisation of Simpson's paradox in terms of the Odds Ratio.

Theorem 3. *Let* T_1 *and* T_2 *be two* 2×2 *contingency tables. Simpson's paradox for* T_1 *and* T_2 *is logically equivalent to the following condition.*

$$\begin{gathered} ((\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1) \wedge \neg (\kappa(T_1 + T_2) > 1)) \\ \vee \\ ((\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1) \wedge \neg (\kappa(T_1 + T_2) < 1)) . \end{gathered} \tag{2}$$

Based on Theorem 3, we state the following result which provides a characterisation of the condition under which SP does not hold.

Theorem 4. *Let* T_1 *and* T_2 *be two* 2×2 *contingency tables. Then Simpson's paradox does not hold for* T_1 *and* T_2 *if and only if*

$$(\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1) \quad \textit{implies} \quad \kappa(T_1 + T_2) > 1, \tag{3}$$

$$\textit{and}$$

$$(\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1) \quad \textit{implies} \quad \kappa(T_1 + T_2) < 1. \tag{4}$$

Homogeneous subpopulations: Mittal (1991) starts the paper with the following sentence: "Homogeneity of subpopulations is a very hard concept to define." She goes on to explain that inspite of the difficulty, the concept is required in statistical studies and that the appropriate concept of homogeneity depends on the context. The paper defines a conception of homogeneity of subpopulations which is shown to be a sufficient condition for SP not to occur. We explain the definition of homogeneity introduced in Mittal (1991) with respect to the 2×2 contingency tables in Definition 1.

Definition 4 (Mittal (1991))**.** *Let* (T_1, T_2) *be as defined in Definition 1. Then* (T_1, T_2) *is said to be homogeneous if one of the following conditions hold.*

$$\begin{aligned} &\max(a_1/b_1, a_2/b_2) < \min(c_1/d_1, c_2/d_2), \qquad \max(c_1/d_1, c_2/d_2) < \min(a_1/b_1, a_2/b_2) \\ &\max(a_1/c_1, a_2/c_2) < \min(b_1/d_1, b_2/d_2), \qquad \max(b_1/d_1, b_2/d_2) < \min(a_1/c_2, a_2/c_2). \end{aligned}$$

Mittal (1991) uses $\leq$ in the above definition to correspond to her definition of SP, i.e., M91. We have changed this to $<$, to correspond to our definition of SP as stated in Definition 2. The reader may refer to Section 2 for a discussion on the relation between M91 and Definition 2.

The following result is a reformulation of the result from Mittal (1991) showing that the notion of homogeneity captured in Definition 4 is sufficient to rule out SP. The proof provided in Mittal (1991) is geometric. In the appendix, we provide a simple proof based on the Odds Ratio characterisation of SP.

Theorem 5. *Let T_1 and T_2 be two 2×2 contingency tables satisfying Definition 4. Then Simpson's paradox does not hold for T_1 and T_2.*

Good and Mittal (1987) define a measure of assocation α for the table T as a function $\alpha(T)$ of T satisfying certain desirable properties. We do not provide the exact properties as these are not directly relevant for the present work. We only note that they showed that the Odds Ratio satisfies these properties and is consequently a measure of association according to their definition of the concept. Given a measure of association α, Good and Mittal (1987) define a notion of homogeneity. The two subpopulations T_1 and T_2 are homogeneous with respect to α if $\alpha(T_1) = \alpha(T_2) = \alpha(T_1 + T_2)$. Applying the definition of homogeneity used by Good and Mittal (1987) to the Odds Ratio κ, the notion of Odds Ratio homogeneity is obtained to be the following.

Definition 5 (Odds Ratio Homogeneity (ORH))**.** *The two subpopulations T_1 and T_2 are homogeneous with respect to the Odds Ratio κ if*

$$\kappa(T_1) = \kappa(T_2) \quad = \quad \kappa(T_1 + T_2). \tag{5}$$

We introduce a weaker form of ORH.

Definition 6 (Weak Odds Ratio Homogeneity (WORH))**.** *The two subpopulations T_1 and T_2 are weakly homogeneous with respect to the Odds Ratio κ if*

$$\kappa(T_1) = \kappa(T_1 + T_2) \quad \textit{or} \quad \kappa(T_2) = \kappa(T_1 + T_2). \tag{6}$$

The next result provides a sufficient condition in terms of WORH under which Simpson's paradox is guaranteed not to hold.

Theorem 6. *Let T_1 and T_2 be two 2×2 contingency tables. If WORH holds for T_1 and T_2, then Simpson's paradox does not hold for T_1 and T_2.*

The notion of positive association of true/false valued random variables X and Y is defined in the following manner (Lindley and Novick (1981), Mittal (1991)).

Definition 7 (Positive Association)**.** *The random variables X and Y are said to be positively associated if* $\Pr[Y|X] > \Pr[Y|\overline{X}]$.

In other words, the conditional probability of Y being true given that X is true is greater than the conditional probability of Y being true given that X is false. So, if we consider two universes (or, sample spaces), one where X is true and the other where X is false, then the probability that Y is true is greater in the first universe than in the second.

If X and Y are positively associated, this is denoted as $X \sim Y$. Note that as stated, the definition of positive association is asymmetric with respect to the variables X and Y, i.e., it is not immediately clear that $\Pr[Y|X] > \Pr[Y|\overline{X}]$ is equivalent to $\Pr[X|Y] > \Pr[X|\overline{Y}]$. So, *a priori* it is not clear that the relation $\sim$ is symmetric[1]. The next result characterises the notion of positive association in terms of Odds Ratio.

Theorem 7. *The random variables X and Y are positively associated if and only if $\kappa(T) > 1$. Similarly, X and $\overline{Y}$ are positively associated if and only if $\kappa(T) < 1$.*

The proof of the above result for positive association of X and Y shows that $\Pr[Y|X] > \Pr[Y|\overline{X}]$ holds if and only if $\kappa(T) > 1$. Similarly, one may show that $\Pr[X|Y] > \Pr[X|\overline{Y}]$ holds if and only if $\kappa(T) > 1$. So, even though the definition of positive association is asymmetric in terms of X and Y, the relation $X \sim Y$ is actually symmetric. Seen in terms of Odds Ratio, the symmetric property of $\sim$ becomes immediate. Showing this property required a somewhat more complicated proof in Mittal (1991). Note that the condition X is positively associated with $\overline{Y}$ is not the same as the negation of the condition that X is positively associated with Y since the possibility $\kappa(T) = 1$ is missing from the characterisations of both the conditions.

Given a pair of contingency tables T_1 and T_2 as in Definition 1 and the associated random variables as described in the discussion following the definition, if X is positively associated with Y in table T_1 (resp. T_2), we denote this as $(X \sim Y)|M$ (resp. $(X \sim Y)|\overline{M}$). The notation $(X \sim Y)|M$ can be understood as saying that X is positively associated with Y with respect to the sub-population represented by T_1, and similarly the notation $(X \sim Y)|\overline{M}$ can be understood as saying that X is positively associated with Y with respect to the sub-population represented by T_2. In particular, X and Y being positively associated with respect to one of the sub-populations represented by T_1 or T_2 does not necessarily imply that X and Y are also positively associated with the other sub-population or, with the overall population. We write $X \sim Y$ if the positive association condition holds for the merged table $T_1 + T_2$.

Using Theorem 7, $(X \sim Y)|M$ (resp. $(X \sim Y)|\overline{M}$) holds if and only if $\kappa(T_1) > 1$ (resp. $\kappa(T_2) > 1$); and $X \sim Y$ if and only if $\kappa(T_1 + T_2) > 1$.

Theorem 3 provides a condition in terms of Odds Ratio which is equivalent to the definition of Simpson's paradox. The connection of positive association to Odds Ratio can be used to translate this condition in terms of positive association of the random variables X and Y. Using Theorems 3 and 7, SP holds for (T_1, T_2) if and only if either $((X \sim Y)|M$ and $(X \sim Y)|\overline{M}$ and $\neg(X \sim Y))$, or $((X \sim \overline{Y})|M$ and $(X \sim \overline{Y})|\overline{M}$ and $\neg(X \sim \overline{Y}))$. In other words, we have the following result.

Theorem 8. *SP holds if and only if* either *X is positively associated with Y in both the sub-populations and not so in the aggregate population,* or *X is positively associated with $\overline{Y}$ in both the sub-populations and not so in the aggregate population.*

The unconditional probabilities of the events $M \wedge Y$, $M \wedge \overline{Y}$, $\overline{M} \wedge Y$, $\overline{M} \wedge \overline{Y}$ and $M \wedge X$, $M \wedge \overline{X}$, $\overline{M} \wedge X$, $\overline{M} \wedge \overline{X}$ are obtained respectively from the following 2×2 contingency tables S_1 and S_2 given below. Note that S_1 and S_2 are obtained from the tables T_1 and T_2 in Definition 1 by aggregating in a different manner. In both S_1 and S_2, the variables M and $\overline{M}$ are associated with the first and second rows respectively. In S_1, Y and $\overline{Y}$ are associated with the first and second columns respectively, while in S_2, X and $\overline{X}$ are associated with the first and second columns respectively.

$$S_1 \quad = \quad \begin{array}{|c|c|}\hline a_1 + c_1 & b_1 + d_1 \\ \hline a_2 + c_2 & b_2 + d_2 \\ \hline \end{array}, \qquad S_2 \quad = \quad \begin{array}{|c|c|}\hline a_1 + b_1 & c_1 + d_1 \\ \hline a_2 + b_2 & c_2 + d_2 \\ \hline \end{array}.$$

[1]Mittal (1991) states that if $X \sim Y$ implies $Y \sim X$, then $\sim$ is reflexive; a relation satisfying the stated condition is usually called symmetric; a relation $\sim$ is said to be reflexive if $X \sim X$ for all X.

Using Theorem 7, $M \sim Y$ (resp. $M \sim X$) if and only if $\kappa(S_1) > 1$ (resp. $\kappa(S_2) > 1$); and $M \sim \overline{Y}$ (resp. $M \sim \overline{X}$) if and only if $\kappa(S_1) < 1$ (resp. $\kappa(S_2) < 1$).

Lindley and Novick (1981) had proposed a necessary condition for SP. The result of Lindley and Novick (1981) was later expanded by Mittal (1991). Mittal (1991) had pointed out a few problematic issues in the arguments used by Lindley and Novick (1981) and had provided a new proof. The proof given in Mittal (1991) itself is essentially a sketch and omits some important details. Also, there are some inaccuracies in the handling of the inequalities. To resolve these issues and for the sake of completeness, we provide a detailed proof of the result in the appendix.

Theorem 9 (Lindley and Novick (1981), Mittal (1991))**.** *Let T_1 and T_2 be a pair of contingency tables and random variables X, Y and M are associated with T_1 and T_2 as stated above. If $(X \sim Y)|M$ and $(X \sim Y)|\overline{M}$ and SP occurs for (T_1, T_2), then either (a) $M \sim X$ and $M \sim Y$; or (b) $M \sim \overline{X}$ and $M \sim \overline{Y}$. Further, the result also holds with Y replaced by $\overline{Y}$.*

4 Simpson's Paradox and Progressive Accumulation of Data

Statistics has been viewed as a method of inductive inference. We mention some examples from the literature where such a connection has been pointed out by leading statisticians. Mahalanobis (1950) wrote: "The importance of statistics in the field of science is due to its supplying the general method for inductive inference." Fisher (1955) had written about the relation between statistical methods and scientific induction. The preface of the book by Rao (1997) suggests that statistics provides a method of codifying inductive reasoning. Mayo and Cox (2006) explore the role of frequentist statistics as a theory of inductive inference.

It is well known that inductive inference is non-monotonic (see for example Levi (2005)). In the present context, the pair of 2×2 contingency tables can be considered to be data. An inference based on such a pair of tables is essentially an inductive inference. Due to the non-monotonic nature of inductive inference, an inference drawn from a particular pair of tables may get invalidated as more data are obtained. One may consider a scenario where data are progressively accumulated over time. Inferences are made at certain time points based on the current data that are available. This leads to the theoretical possibility that the data indicates SP at a certain point of time which disappears as more data accumulates and then reappears with further accumulation of data. We present a result to show that such a theoretical possibility is indeed possible and the appearance and disappearance of SP can continue ad infinitum.

Let $\mathcal{T} = (T_1^{(k)}, T_2^{(k)})_{k \geq 0}$ be a sequence of pairs of 2×2 contingency tables, where

$$T_1^{(k)} = \begin{array}{|c|c|} \hline a_1^{(k)} & b_1^{(k)} \\ \hline c_1^{(k)} & d_1^{(k)} \\ \hline \end{array}, \quad T_2^{(k)} = \begin{array}{|c|c|} \hline a_2^{(k)} & b_2^{(k)} \\ \hline c_2^{(k)} & d_2^{(k)} \\ \hline \end{array}.$$

We define $\mathcal{T}$ to be monotonic if for all $k \geq 0$,

$$\begin{gathered} a_1^{(k+1)} \geq a_1^{(k)}, \quad b_1^{(k+1)} \geq b_1^{(k)}, \quad c_1^{(k+1)} \geq c_1^{(k)}, \quad d_1^{(k+1)} \geq d_1^{(k)}, \text{ and} \\ a_2^{(k+1)} \geq a_2^{(k)}, \quad b_2^{(k+1)} \geq b_2^{(k)}, \quad c_2^{(k+1)} \geq c_2^{(k)}, \quad d_2^{(k+1)} \geq d_2^{(k)}. \end{gathered}$$

In other words, the pair of contingency tables $(T_1^{(k+1)}, T_2^{(k+1)})$ is obtained from the pair $(T_1^{(k)}, T_2^{(k)})$ by accumulating additional data. From an empirical data perspective, some initial data collection leads to the pair of tables $(T_1^{(0)}, T_2^{(0)})$. Obtaining additional data over and above that which has been already accounted for in the pair $(T_1^{(0)}, T_2^{(0)})$ leads to the pair of tables $(T_1^{(1)}, T_2^{(1)})$, further data successively leads to the pairs of tables $(T_1^{(2)}, T_2^{(2)})$, $(T_1^{(3)}, T_2^{(3)})$, and so on.

Consider Figure 1. Case 3 represents $(>, >, <)$, i.e., the '$>$' relation holds for both T_1 and T_2, but '$<$' holds for $T_1 + T_2$. Similarly, Case 25 represents $(<, <, >)$. Both Cases 3 and 25 are instances of SP. For the sake of convenience, let us call Case 3 to be positive SP and Case 25 to be negative SP. Similary, Case 1 represents $(>, >, >)$ and we will call this to be positive alignment and Case 27 represents $(<, <, <)$ which we call to be negative alignment.

Theorem 10. *There exists a monotonic sequence $\mathcal{T} = (T_1^{(k)}, T_2^{(k)})_{k\geq 0}$ such that the following holds.*

1. $(T_1^{(0)}, T_2^{(0)})$, $(T_1^{(4)}, T_2^{(4)})$, $(T_1^{(8)}, T_2^{(8)}), \ldots$ *are instances of positive SP.*
2. $(T_1^{(1)}, T_2^{(1)})$, $(T_1^{(5)}, T_2^{(5)})$, $(T_1^{(9)}, T_2^{(9)}), \ldots$ *are instances of positive alignment.*
3. $(T_1^{(2)}, T_2^{(2)})$, $(T_1^{(6)}, T_2^{(6)})$, $(T_1^{(10)}, T_2^{(10)}), \ldots$ *are instances of negative SP.*
4. $(T_1^{(3)}, T_2^{(3)})$, $(T_1^{(7)}, T_2^{(7)})$, $(T_1^{(11)}, T_2^{(11)}), \ldots$ *are instances of negative alignment.*

Theorem 10 shows that it is possible to have progressively accumulated data where positive SP and negative SP are interleaved with positive and negative alignments and such togglings repeat ad infinitum. Consider $(T_1^{(0)}, T_2^{(0)})$. Since this is an instance of positive SP, an empirical scientist may invoke some resolution mechanism and draw a conclusion. If the conclusion is to accept the relation suggested by the sub-populations, then further data as in $(T_1^{(1)}, T_2^{(1)})$ will confirm such a conclusion, while if the conclusion is to accept the relation suggested by the aggregate population, then $(T_1^{(1)}, T_2^{(1)})$ will invalidate the conclusion. Further data will exhibit negative SP followed by negative alignment and then again positive SP and so on. This suggests that any explanation of Simpson's paradox which aims to be comprehensive should keep in focus the inductive nature of inference from data.

5 Examples of Simpson's Paradox

The literature provides several examples of SP. In this section, we discuss some of these examples. There are two distinct aspects to the examples of SP that appear in the literature. The first aspect is numerical which pertains to the actual data available in a pair of 2×2 contingency tables, while the second aspect is the context in which the contingency tables are discussed. In our discussion, we separate the numerical aspect from the contextual aspect.

Pearl (2014) had commented that Lindley and Novick (1981) had raised "Simpson's paradox to new heights" by providing two contexts for the same data. While Lindley and Novick's work was indeed influential, two separate contexts for the same data appear in the much earlier paper by Simpson (1951).

We describe the context in terms of the true/false valued random variables X, Y and M identified in Section 2. To refresh memory, we mention that M and $\overline{M}$ correspond to the tables T_1 and T_2 respectively; X and $\overline{X}$ correspond to the first and the second rows respectively; and, Y and $\overline{Y}$ correspond to the first and second columns respectively.

Example 1 (Simpson 1951):

$$T_1 = \begin{array}{|c|c|}\hline 4 & 3 \\ \hline 8 & 5 \\ \hline\end{array}, \quad T_2 = \begin{array}{|c|c|}\hline 2 & 3 \\ \hline 12 & 15 \\ \hline\end{array}, \quad T_1 + T_2 = \begin{array}{|c|c|}\hline 6 & 6 \\ \hline 20 & 20 \\ \hline\end{array}.$$

For this example, we have $A_1 < C_1$, $A_2 < C_2$ and $\mu = \nu$. So, the association which is visible in the sub-populations ceases (as opposed to being reversed) upon aggregation.

Context: Cards (Simpson 1951). The context concerns a deck of cards. M and $\overline{M}$ correspond to dirty and clean cards respectively; X and $\overline{X}$ correspond to court (King, Queen, Knave) cards and plain cards respectively; Y and $\overline{Y}$ correspond to red and black cards respectively. The goal is to examine whether the proportion of court cards is associated with colour. It would appear to be so from the individual tables T_1 and T_2, but such association disappears from the aggregated table. As Simpson noted, in this context, it is the aggregated table which makes sense.

Context: Treatment (Simpson 1951). The context concerns whether patients are alive or dead after receiving treatment or not. M and $\overline{M}$ correspond to male and female patients respectively; X and $\overline{X}$ correspond to untreated and treated patients respectively; Y and $\overline{Y}$ correspond to being alive and dead respectively. So, a positive association between treatment and survival is seen separately for male and female population groups, but no such association is observed from the aggregated data. Simpson raises the question of "sensible" interpretation noting that a treatment which is beneficial to both males and females cannot be rejected as being useless for the race. The wording of the question suggests that Simpson considers the inference from the disaggregated tables to be more sensible, but does not actually state this.

Example 2 (Blyth 1971):

$$T_1 = \begin{array}{|c|c|} \hline 1000 & 9000 \\ \hline 50 & 950 \\ \hline \end{array}, \quad T_2 = \begin{array}{|c|c|} \hline 95 & 5 \\ \hline 5000 & 5000 \\ \hline \end{array}, \quad T_1 + T_2 = \begin{array}{|c|c|} \hline 1095 & 9005 \\ \hline 5050 & 5950 \\ \hline \end{array}.$$

For this example, we have $A_1 > C_1$, $A_2 > C_2$ and $\mu < \nu$. So, the association which is visible in the sub-populations is reversed upon aggregation.

Context: Treatment (Blyth 1971). The context concerns whether local or outstation patients are treated. M and $\overline{M}$ correspond to local and outstation patients respectively; X and $\overline{X}$ correspond to a new treatment and the standard treatment respectively; Y and $\overline{Y}$ correspond to being alive and dead respectively. Blyth explains the reversal by suggesting that local patients are much less likely to recover and the treatment was given mostly to local patients. A treatment will show a poor recovery rate if it is given mostly to the most seriously ill patients.

Example 3 (Gardner 1976):

$$T_1 = \begin{array}{|c|c|} \hline 5 & 6 \\ \hline 3 & 4 \\ \hline \end{array}, \quad T_2 = \begin{array}{|c|c|} \hline 6 & 3 \\ \hline 9 & 5 \\ \hline \end{array}, \quad T_1 + T_2 = \begin{array}{|c|c|} \hline 11 & 9 \\ \hline 12 & 9 \\ \hline \end{array}.$$

For this example, we have $A_1 > C_1$, $A_2 > C_2$ and $\mu < \nu$. So, the association which is visible in the sub-populations is reversed upon aggregation.

Context: Poker Chips (Gardner 1976). The context concerns different coloured poker chips placed in different coloured hats. M and $\overline{M}$ correspond to two tables (i.e., pieces of furniture, not tabulation of data); X and $\overline{X}$ correspond to black and grey hats respectively; Y and $\overline{Y}$ correspond to being blue and white chips respectively.

Example 4 (Lindley and Novick 1981):

$$T_1 = \begin{array}{|c|c|}\hline 18 & 12 \\ \hline 7 & 3 \\ \hline\end{array}, \quad T_2 = \begin{array}{|c|c|}\hline 2 & 8 \\ \hline 9 & 21 \\ \hline\end{array}, \quad T_1 + T_2 = \begin{array}{|c|c|}\hline 20 & 20 \\ \hline 16 & 24 \\ \hline\end{array}.$$

For this example, we have $A_1 < C_1$, $A_2 < C_2$ and $\mu > \nu$. So, the association which is visible in the sub-populations is reversed upon aggregation.

Context: Treatment (Lindley and Novick 1981). The context concerns treatments to patients. M and $\overline{M}$ correspond to male and female patients respectively; X and $\overline{X}$ correspond to being treated and not treated respectively; Y and $\overline{Y}$ correspond to recovery or not respectively. In this context, not being treated appears to be better individually for both male and female and patients, but upon aggregation, it appears that treatment is better. The explanation forwarded by Lindley and Novick is that males have been mostly assigned to the treatment group and females to the control group. They also suggested that perhaps the doctor distrusted the treatment and was reluctant to give it to females where the recovery rate is much lower. For this context, the disaggregated data makes more sense, since it would be better not to recommend the treatment when it is clearly inferior for both males and females. Lindley and Novick had further suggested that instead of male/female distinction, the groupings can be considered based on some factor which may be difficult to determine such as a genetic classification.

Context: Agriculture (Lindley and Novick 1981). The context concerns coloured varities of plants which are short or tall and have low or high yield. M and $\overline{M}$ correspond to tall and short plants respectively; X and $\overline{X}$ correspond to white and black varieties respectively; Y and $\overline{Y}$ correspond to high and low yield respectively. For this context, the aggregated data makes more sense, since one would choose the white variety which provides higher yield overall ignoring the individual yield information provided by the tall and the short varities.

Example 5 (Hand 1994):

$$T_1 = \begin{array}{|c|c|}\hline 255 & 174 \\ \hline 156 & 102 \\ \hline\end{array}, \quad T_2 = \begin{array}{|c|c|}\hline 88 & 222 \\ \hline 82 & 175 \\ \hline\end{array}, \quad T_1 + T_2 = \begin{array}{|c|c|}\hline 343 & 396 \\ \hline 238 & 277 \\ \hline\end{array}.$$

For this example, we have $A_1 < C_1$, $A_2 < C_2$ and $\mu > \nu$. So, the association which is visible in the sub-populations is reversed upon aggregation.

Context: Psychiatry (Hand 1994). The context concerns the proportion of male and female patients of different age groups in psychiatry wards in two different years. M and $\overline{M}$ correspond to patients having ages ≤ 65 and those having ages > 65 respectively; X and $\overline{X}$ correspond to numbers of patients in the years 1970 and 1975 respectively; Y and $\overline{Y}$ correspond to male and female respectively. Hand contends that both the disaggregated and the aggregated table provide meaningful answers, depending on what question one asks. If one wishes to know whether the proportion of males increase, then the aggregated table provides the answer, while, if one wishes to know for patients of a given age, whether the proportion of males increase, then the disaggregated tables provide the answer.

Our deconstruction of the examples of SP into the numerical and the contextual aspect raises interesting possibilities. One may take the numerical aspect of one example and match with the contextual example of another example. This gives rise to a new example of SP. One may then contrast the explanation of the context provided for the original numerical part with that of the new numerical part. To illustrate this procedure, consider combining the Lindley and Novick (1981) treatment context with the numerical example of Blyth (1971). One then notes that recovery rate for treated patients are higher than those for untreated patients in both male and female groups, but overall the recovery rate is lower for treated patients. Should one then go by the individual tables and recommend the treatment? Since the recovery rate for the aggregated data is lower for treatment, can one simply ignore this finding without further exploration of possible confounding factors? In the medical context, it is perhaps better to err on the side of safety and not recommend the treatment. In contrast to these doubts, the Lindley-Novick treatment context combined with the Lindley-Novick's numerical example, led to the suggestion of going by disaggregated tables and this was considered reasonable. The point here is that an explanation or resolution of SP is for the holistic combination of data-plus-context. Considering an explanation to be only for the context while ignoring the data may not be satisfying.

6 The Surprise Element

The reversal property in a pair of contingency table is called a paradox, since it appears to be surprising. Following the distinction in Section 5, we consider the numerical and the contextual aspect to be separate. There is nothing surprising or unsurprising about the numerical aspect. To highlight that the numerical aspect is not surprising, we mention the connection of SP to determinants of 2×2 matrices. Consider the tables T_1 and T_2 to be 2×2 matrices. The condition $A_1 > C_1$, $A_2 > C_2$ and $\mu < \nu$ is equivalent to the condition $\det(T_1) > 0$, $\det(T_2) > 0$ and $\det(T_1 + T_2) < 0$, where $\det(T)$ denotes the determinant of T. (This is easily seen through the connection to the Odds Ratio characterisation given in Theorem 3.) There is nothing surprising in the fact that the determinants of T_1 and T_2 are both positive while the determinant of $T_1 + T_2$ is negative.

In view of the above, the surprise element arises from the context. We provide an explanation of why such a reversal appears to be surprising. Human intuition is trained to expect *uniformity* in various aspects of reasoning. For example, in elementary deductive logic, suppose that a statement P_1 implies a statement Q and another statement P_2 also implies Q. The statements P_1 and P_2 provide support to the inference of Q. Aggregation of the supports will mean $P_1 \vee P_2$ and then we still have $P_1 \vee P_2$ implies Q. More generally, human intuition expects that if some property holds for parts, then if the parts are put together the same property will still continue to hold for the whole.

Such uniformity, on the other hand, does not necessarily hold in all possible scenarios. Whenever parts behave differently from the whole, i.e., all the parts display a particular behaviour while the whole does not display such a behaviour or, displays a behaviour which is in contrast to the behaviour of the parts, a surprise element is involved and this gives rise to a paradox. Simpson's paradox is clearly of this kind. A property which holds for the two sub-populations ceases to hold or reverses when the populations are combined. This can be seen as falling within the broad class of aggregation paradoxes. There are some well known examples of such paradoxes where a feature which holds for individuals reverses itself upon aggregation. As examples, we refer to the Ostrogorski paradox (Daudt and Rae (1976)) and the discursive dilemma (Kornhauser and Sager (1986), List and Petit (2002)) where choices/judgements made at the individual level are not preserved upon aggregation. Another famous paradox involving parts and whole is the Banach-Tarski paradox, which shows that a unit sphere can be divided into finitely many parts and assembled together to obtain two unit spheres; the paradox being explained by the fact that the individual parts are defined in a manner that precludes these parts from having a volume, whereas the whole sphere has a well-defined volume.

While explaining Simpson's paradox from the viewpoint of causal graphs, Pearl (2014) comments that

> "... it is hard, if not impossible, to explain the surprise part of Simpson's reversal without postulating that human intuition is governed by causal calculus together with a persistent tendency to attribute causal interpretation to statistical associations."

We consider two objections to the above statement. The first objection is that there is indeed a way to explain the surprise element without assuming that causal considerations are woven into human intuition. The second objection is that there are examples of SP in the literature which do not readily admit a causal modelling. The effect of the second objection is to show the incompleteness of the causal approach in explaining the surprise element of SP while the first objection shows a different method of explaining this surprise element.

Regarding the first objection, we note that as argued above, there is a different and perhaps more fundamental perspective which explains the surprise element of SP, i.e., human intuition expects a kind of uniformity where the whole behaves in a manner similar to the parts. Deviations create surprise and lead to other paradoxes as mentioned above. As for the second objection, in contrast to what Pearl suggests, there are examples of SP where there is no natural interpretation using causal models. One such example, is the poker chip context of Gardner (1976) which has been described in Section 5. In this example, there are no natural causal issues. It is absurd to say that the colours of the hats cause the colours of the chips or vice versa; similarly, it is absurd to say that the physical tables upon which the hats have been placed cause either the colours of the hats or the colours of the chips. Since a causal diagram is a model, one may of course, draw any model that one wishes, but there would be no way to determine that one particular model is to be preferred to another. A scenario where all models are equally (un)likely hardly provides a useful theory. In effect, viewing this example using causal models is like trying to fit a square peg into a round hole.

Another example which does not admit a causal model has been described by Bandyopadhyay et al. (2011). In this example, there are two bags with each bag containing marbles which are either big or small and are either red or blue. Though a numerical example is not provided, it is possible to construct an example of a pair of 2×2 contingency tables corresponding to the bags and rows/columns corresponding to size/colour of marbles which exhibit SP. Again, there are no natural causal issues in this example: bags do not cause size/colour of marbles; size does not cause colour and vice versa.

The above examples show that viewing SP through the prism of causal models potentially leaves out examples of SP. Note that our explanation of the surprise element arising from the expectation of uniformity applies to both the above examples.

7 Causal Structures and SP: Some Questions

Consider a set of random variables V. Based on prior knowledge, the variables are ordered by hypothesised cause-effect relations. The variables and the cause-effect relations can be conveniently described using a directed acyclic graph (DAG), where the nodes are the variables and there is an arrow from one node to another if the first node is a cause of the second. The joint probability distribution of the random variables in V is denoted by $P(v)$. A causal structure consists of the DAG and the joint probability distribution. Such a structure was introduced by Pearl (1995, 2009). For a random variable X, and a value x in the domain of X, the event $X = x$ is well defined and one may consider conditional probabilities of the form $\Pr[Y = y|X = x]$, where Y is another random variable in V and y is a value in the domain of Y. Pearl (2009) defines an important concept called "do", where $\text{do}(x)$ corresponds to setting $X = x$ everywhere and removing the variable X from the DAG. This alters the structure of the DAG and the joint probability distribution. One can consider the probability of the event $Y = y$ in the altered causal structure and this probability is denoted as $\Pr[Y = y|\mathrm{do}(x)]$. In Pearl (1995) this notion was termed intervention. An intervention is not always possible in practice. A remarkable result proved in Pearl (1995) shows that under certain conditions the effect of intervention can be determined from the causal structure without actually performing the intervention.

Pearl (1995, 2009) defines an important concept related to causal graphs, namely the back-door criteria. Given a pair of variables (X, Y), another variable Z (or, possibly a set of variables) satisfies the back-door criterion relative to (X, Y), if Z is not a direct descendant of X, and Z blocks every path between X and Y which contains an arrow into X. The back-door criterion captures the idea that the variable Z blocks "spurious" paths entering X. A corresponding notion of front-door criterion has also been defined.

While SP was not considered in Pearl (1995), later work by Pearl (2009, 2014) applied the causal structure described above and the notion of back-door criterion to the analysis of SP. Certain graph structures are stated to be incapable of exhibiting SP which in effect constitute sufficient conditions for not SP. Such sufficient conditions are different in nature from the sufficient conditions for not SP which have been derived in Section 3. The back-door criterion has been put forward by Pearl as a resolution mechanism for SP. Simple rules are provided for determining whether to accept the conclusion of the individual data, or the aggregated data, or, none.

The causal modelling and the subsequent analysis of SP is a remarkable achievement. Nonetheless, there are some relevant questions regarding this approach which do not seem to have been adequately addressed in the literature. Below we list down some such questions.

Can all examples of SP be modelled using causal structures? In Section 6, we have provided two scenarios where there is no natural associated causal model. This shows that not all examples of SP have a natural interpretation in terms of causal structures.

Can all examples of SP be resolved using causal structures? Since all examples of SP cannot be modelled using causal structures, it clearly follows that all examples of SP cannot be resolved using such structures. For examples of SP which do admit causal modelling, there is an implicit claim in Pearl (2014) that a resolution can be obtained within the causal model. There does not, however, seem to be a proof of such a claim and it is not even clear whether such a claim can be formally stated and proved.

Use of deductive methods to resolve an inductive inference. It has been argued in Section 4 that statistical inference is in general a method of inductive inference. A causal structure on the other hand, assumes that the causal relations are known before hand. This raises the question of how one arrives at these causal relations? Such relations may be considered to have been provided by domain experts. This, however, deflects the question to another level, since one could pose the question as to how the domain experts would have arrived at these causal relations. From a scientific point of view, we see no way of justifying the determination of causal relations except by referring to previous observations/data. But, then one would have to look for justifications of the procedures for deriving the causal relations from data, leading to a circularity in the argument. An argument along the same line has been given earlier in Freedman (1995) who remarked that "Validation of causal models remains an unsolved problem" and also "... causal laws, ..., are assumed rather than inferred from the data."

At several places, Pearl has critiqued statisticians for avoiding terminology based on causes. For a discussion on use of ‘correlation’ by Pearson and Yule, we refer to Aldrich (1995). The paper also identifies several aspects of Pearson’s thoughts on causality and remarks that Pearson’s “complete position was that causation *is* correlation, or more precisely the limiting case of correlation, *except* when the correlation is spurious, when correlation is not causation.” (Emphasis as in the original.) It is perhaps not surprising that the field of statistical inference will be careful in talking about causation. After all, statistical inference is a form of inductive inference and the skepticism to inferring causes using inductive inference has a long history.

A causal model incorporates known relationships among the variables. How does one determine whether the known relationships are *all* the relations that hold? Along this line of reasoning, it has been observed by Cox and Wermuth (1995) that the “back-door criterion requires there to be no ‘common cause’ ... that is not blocked out by the observed variable” and that it is “precisely doubt about such assumptions that makes epidemiologists, for example, wisely in our view, so cautious in distinguishing risk factor from causal effects.” One can never really be sure that a given causal model represents the entire cause-effect relations that hold among the given variables. In this context, we mention that the question of how to ensure that all hidden factors have been eliminated was one of the objections raised by the ancient Cārvāka school of thought to inductive inference (see Page 15 of Chakrabarti (2010)).

The modelling of causality using causal structures is a brilliant piece of deductive mathematical machinery. The *assumed* DAG and the probability distribution form the axioms of the deductive theory (along with the usual axioms of logic and probability). A theorem in the theory of causal structures is a statement about the DAG and the probability distribution and is proved using usual deductive mathematical methods. So, the remarkable theorem that it is possible to derive the effects of intervention without actual experiments is after all a deductive result. As such, it is non-ampliative in the sense that it does not convey more knowledge than what was already contained in the design of the causal structure. Statistical inference, on the other hand, is ampliative. So, at a fundamental level the applicability of modelling using causal structures to statistical inference is limited.

As has been argued above, statistical inference is one form of inductive inference. In fact, as quoted above, Mahalanobis considers statistics to be *the* general method of inductive inference. As a method of inductive inference, broadly speaking one may consider one of the goals of statistical inference to derive cause-effect relations. Statistical inference, however, avoids talking about cause-effect relations since statistical techniques cannot determine such relations. On the other hand, causal structures modelling starts out by postulating causal relations and the associated probability distributions. So, considering statistical inference to be an exercise in causal modelling is like postulating what one wishes to derive. In this context, it may be noted that Russell had raised objections about “introducing entities with implicit definitions, that is, as being those things that obey certain axioms or ‘postulates’.” (Linsky (2019))

An important aspect of any inductive inference is non-monotonicity. In the context of SP, this issue has been discussed in Section 4. We note that Imbens and Rubin (1995) remark that “the monotonicity assumption is difficult to represent in a graphical model”.

8 To Aggregate or Not to Aggregate?

In Section 3, we have provided several sufficient conditions for SP not to hold. These are conditions on the sub-populations. If any of these conditions hold, then the paradox will not arise. Among these sufficient conditions are two homogeneity conditions, one of which is a slight modification of a condition due to Mittal (1991) and the other is a new condition, namely the WORH condition. The sufficient conditions in general and the homogeneity conditions in particular, are not necessary. Consequently, the question arises as to whether one should accept the conclusion of the aggregated table when the homogeneity conditions do not hold. If one takes the position that the data are acceptable only if the homogeneity condition holds, then there will not be any paradox and effectively this would make the resolution of the paradox very simple. However, if a homogeneity condition does not hold, then this does not necessarily mean that aggregation should not be done. There are scenarios, where the homogeneity conditions do not hold, yet, it is the aggregated table which is the correct answer. From Section 5, one may note such examples arising in Simpson (1951) and the agricultural example in Lindley and Novick (1981). Nevertheless, the non-satisfaction of the homogeneity conditions does raise a red flag. In this context, we recall the caveat from Mittal (1991) who recommended that if the homogeneity condition she identified is not satisfied, then "a closer look at the data is warranted before amalgamation ... Such a look may reveal some hidden factors in the data that might make amalgamation unwise."

Simpson's paradox has been investigated from the empirical perspective by Spanos (2020) in the parametric framework using the notion of statistical mis-specification. Homogeneity considerations arise in this investigation and Spanos (2020) points out that the untrustworthy associations in the Cohen and Nagel (1934) example is due to the fact that "the two populations are not homogenous." Several other examples are also explained by showing that aggregation leads to unreliable models. We note, on the other hand, that Spanos (2020) does not appear to provide an example, where homogeneity conditions are violated and yet the aggregated data provides the correct answer. As mentioned above, such an example is the agricultural data example in Lindley and Novick (1981). It would be of interest to know whether the approach suggested by Spanos (2020) can explain this (and similar) examples.

An explanation of Simpson's paradox from the view point of confirmation theory has been put forward by Fitelson (2017). While the explanation is interesting, the work does not consider the important problem of "what to do?" when faced with an instance of SP. In fact, it would be of interest to perform a detailed analysis of the examples described in Section 5 based on Fitelson's approach.

From the discussion following Definition 1, we have an instance of the paradox if $\Pr[Y|X,M] > \Pr[Y|X,\overline{M}]$, $\Pr[Y|\overline{X},M] > \Pr[Y|\overline{X},\overline{M}]$, but $\Pr[Y|X] < \Pr[Y|\overline{X}]$. One way to understand how this can happen is to use the conditional probability expansions of $\Pr[Y|X]$ and $\Pr[Y|\overline{X}]$ as follows.

$$\begin{aligned}
\Pr[Y|X] &= \Pr[Y|X,M]\Pr[M|X] + \Pr[Y|X,\overline{M}]\Pr[\overline{M}|X], \\
\Pr[Y|\overline{X}] &= \Pr[Y|\overline{X},M]\Pr[M|\overline{X}] + \Pr[Y|\overline{X},\overline{M}]\Pr[\overline{M}|\overline{X}].
\end{aligned}$$

So, both $\Pr[Y|X]$ and $\Pr[Y|\overline{X}]$ can be seen as weighted sums, where the weights for $\Pr[Y|X]$ are $\Pr[M|X]$ and $\Pr[\overline{M}|X]$ and the weights for $\Pr[Y|\overline{X}]$ are $\Pr[M|\overline{X}]$ and $\Pr[\overline{M}|\overline{X}]$. Since the weights corresponding to $\Pr[Y|X,M]$ and $\Pr[Y|\overline{X},M]$ are not equal and similarly, the weights corresponding to $\Pr[Y|X,\overline{M}]$ and $\Pr[Y|\overline{X},\overline{M}]$ are not equal, the inequalities $\Pr[Y|X,M] > \Pr[Y|X,\overline{M}]$ and $\Pr[Y|\overline{X},M] > \Pr[Y|\overline{X},\overline{M}]$ are not preserved on aggregation.

The above provides a simple mathematical explanation of the paradox and has been pointed out by Hand (1994) who also mention that this is an "inadequate resolution." In the words of Hand (1994), "the issue is whether or not ... to ask a conditional question and, in fact, the issue of whether or not to condition is ubiquitous. ... Neither analysis is right and the other wrong – it depends on what we want to find out." The issue of formulating separate questions whose answers are provided by the disaggregated and the aggregated tables is another dimension of the typical 'What to do?' question that an investigator has to contend with when faced with an instance of SP. We may take this view and apply it to some of the examples in Section 5.

1. Consider the example of Gardner (1976). If we wish to maximise the probability of drawing a blue chip, then if the chips are in the disaggregated format, we should draw from the black hats,

while if the chips are in the aggregated format, then we should draw from the grey hat. Since, the chips will be in either the disaggregated or the aggregated format, we have a complete resolution of the paradox in this particular case.

2. Consider the agricultural example of Lindley and Novick (1981). If both short and tall varieties are available in adequate quantity, then it is better to ignore this aspect and plant only the white variety. On the other hand, if only the short variety, or, only the tall variety is available, then clearly one should go by the disaggregated table and plant the black variety. So, depending on the context, both the disaggregated and the aggregated tables provide useful information.

The question arises as to whether the above kind of analysis can be applied to all examples of SP? It seems that there are difficulties in doing so. Consider the Lindley and Novick (1981) example of treatment described in Section 5. In this example, the moot question is whether to apply the treatment or not, to which the answer is provided by the disaggregated tables; it seems difficult to come up with a question whose answer is given by the aggregated table.

In Pearl (2014), it has been mentioned that "in certain models the correct answer may not lie in either the disaggregated or the aggregated data." An explanation for such a situation has been provided in terms of the backdoor criterion. We would like to highlight that there is no known example of SP in the literature where neither the disaggregated nor the aggregated data provide the correct answer. This raises the question of whether this possibility is only a mathematical artifice, or, whether such cases can really arise in practice? Further, we note that the possibility that for a particular example, both the disaggregated and the aggregated data could be correct depending on the context, does not seem to have been considered by Pearl.

In the previous sections, we have raised several questions regarding the claimed resolution of the paradox using the causal approach. We do not, however, wish to end on a negative note. As we have mentioned several times, the causal structure approach is indeed an excellent piece of mathematical machinery. In the context of SP, for the examples where it is applicable, it can certainly provide valuable information in the sense that in many situations hypothesising a certain kind of cause-effect relation can provide suggestions on whether to accept the conclusion of the disaggregated, or, the aggregated data, or, none. One must, however, keep in mind that any such conclusion is contingent on the hypothesised cause-effect relations. Our view is that in resolving SP, one should take into consideration other statistical aspects along with the causal graph analysis.

9 Broader Perspective: Logic, Probability, and Statistics

Issues involving causation in statistics date to more than one hundred year ago. There are at least four possible positions regarding the role and status of causality in statistics. They are (i) radical causal skepticism, (ii) modest causal skepticism, (iii) causal/statistical compatibilism, and (iv) full-blown adherence to causality. Radical skepticism about causation is aptly captured by Karl Pearson for whom correlation is all there is. For him, there is no room for causation in statistics (See Aldrich, 1995 for a nuanced approach to Pearson). Causation skepticism dates to Pearson (1910) when Fisher's theory of randomization had not yet developed. In his words, (1910 quoted in Aldrich, 1995.)[2]

[2]In one sense, causality existed in statistics before 1950's. The role of randomization is where causality arises after Pearson left. It was precisely in the hand of Fisher that the theory was developed. There was an intellectual feud between Fisherians and Wright about causality and the role of path-diagrams in data analysis. It resulted in a sustained neglect of the appreciation of path analysis in the statistical community (see, Pearl and Mackenzie, 2018).

> "It is this conception of correlation between two occurrences embracing all relationships from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation (Pearson, 1910, p.157)."

No wonder that Russell, being an admirer of Pearson's work, shares his causal skepticism in science. Pearl calls Pearson "causality's worst adversary." (Pearl, 2009, P.105.)

A weaker position about statisticians' stance toward causal language can be attributed to Novick and Lindley's influential paper in which they address Simpson's Paradox from a Bayesian statistic. For them, informal causal talk is possible, but, according to them, causation is not a well-defined term to be presented in statistics. They write, "One possibility would be to use the language of causation .... We have not chosen to do this; nor to discuss causation, because the concept, although widely used, does not seem to be well-defined (Novick and Lindley, 1981)." Their view on causation in statistics resembles Earman's oft-quoted expressions about "the wooliness" of causal notion in physics. We call their position, modest causal skepticism. Causal theorists such as Glymour, Meek, and Pearl took them to task for trying hard to avoid causal language when clearly, according to these proponents of causal theories, causal notions are brought in their paper from the backdoor (because of the use of exchangeability). Elsewhere, Pearl has gone further than they and criticized the statistical community with a blistering language (Pearl, 2009, P.177.).

> "Simpson's paradox helps us to appreciate both the agony and the achievement of this generation of statisticians. Driven by healthy 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 (i.e., *causal*) after all." (Emphasis is ours.)

Those who provide a full-blown defense of causality in statistical side are Pearl on the one hand, and Spirtes, Glymour, and Schienes on the other. Pearl is a strong proponent of causality. Pearl is interested in the "Why Question" where a connection needs to be established between a known cause and a known effect for a better grasp of the mechanism at play. He discusses in particular the role causality plays in scientific inference. This discussion is in terms of what he calls "the ladder of causation" consisting of three levels. The first level is the world of association in which statistics plays a role, and there is no causation properly so called. The second level is where intervention plays a role. We intervene in a system to know its causal mechanism with the help of a causal model. The third and final level in the ladder is the world of counterfactuals in which there are no data to offer an insight into counterfactual reasoning. We ask, "What would be the case for y if x had happened?"[3]

Pearl is interested in understanding the foundation of a theory of causality. However, he is keener on developing a mathematical language to furnish the theoretical underpinning of causality. He thinks that statistical apparatus fails to capture this kind of counterfactual reasoning as statistics, according to him, involves and pivots only around data. With the information of an able investigator, we can propose a causal model to understand the mechanism of some phenomenon where data could provide some clue as to their mechanism, but not necessarily the full story of the phenomenon at stake.

The CMU philosophers/statisticians share Pearl's enthusiasm for the significance of causality in contemporary methodology. There is at least one difference between Pearl and the CMU philosophers and statisticians, however. Pearl is not interested in the discovery of a cause, as he builds causal models by drawing on expert knowledge and available data. In contrast, CMU philosophers/statisticians are

[3] David Lewis has contributed considerably to our understanding of counterfactuals (Lewis, 1973.)

interested in causal discovery. To this end, they suggest a subject-matter-neutral automated causal inference engine that provides causal relationships among variables from observational data using information on their probabilistic correlations and assumptions about their causal structure. Both causal accounts are anti-reductionists in the sense that a causal relationship among events cannot be defined in terms of any other concepts in addition to the fact that data cannot provide a full narrative about the causal process. According to them, we are required to have causal assumptions to begin with to make causal inferences.

We reconstructed Pearl's strongest stance toward causality including his critical stance against statistical community for disregarding the need for casual language in providing a theory of scientific inference. However, he has changed his rhetoric and become less brutal about statistics being able to express causal language. Presumably, the most weakened stance he has taken toward statistics is to enrich statistics when it needs causal language as well as employing statistical language to provide a theory of causal inference. We call this position of the co-existence of statistics with causality, the causality-statistics compatibilism.

Pearl writes "causation is not merely an aspect of statistics; it is an addition to statistics, an enrichment that allows statistics to uncover workings of the world that traditional methods cannot." (Pearl et al. 2016.) So, the analysis of what Pearl and his co-authors provide is that investigators need to know how and why causes influence their effects by analyzing data in a study. An analysis of the data helps appreciate why a cause that holds in one context may not hold in another. Presumably Pearl's view is that generalizability across domains is motivated by how science works, and not just from how statistics functions. The title of their book *Causal Inference in Statistics* is suggestive. It captures a mutual feedback process that causal inference can incorporate statistics, and the converse is the case; yet they are distinct disciplines (Pearce and Lawlor, 2016).

To summarize, there are at least four accounts of the role of statistics in causality debate. They are (i) radical causal skepticism, (ii) modest causal skepticism, (iii) causal/statistical compatibilism, and finally, the full-blown causal theory of scientific inference.

Keeping in mind our interest in Simpson's Paradox, we will only mention two extreme stances toward causality. They are the Russell-Pearson position where causality is completely eschewed, and the Pearl-CMU view in which causality is embraced whole-heartedly. Our Simpson's Paradox-centered stance concerning causality does not adopt either of the extreme positions.

The underlying theme of this paper is to contend that we need causal language combined with statistical/probabilistic tools in which philosophy helps sort out among different types of questions. We raised at least three questions regarding the paradox: (i) Why is Simpson's Paradox paradoxical? (ii) What are the conditions for generating the paradox, and (iii) What-to-do when confronted with a Simpson's Paradoxical type situation? We disagreed with the causal resolution regarding the surprising element in the paradox. While discussing different definitions of Simpson's Paradox, we showed that logic/probability coupled with statistical language devoid of causal intuition is adequately powerful to represent the paradox. Therefore, Simpson's Paradox has nothing to do with causal considerations in addressing the first two questions. For the what-to-do question, causal considerations, along with other statistical tools, are important in making inductive inferences. Therefore, we are, in a sense, compatibilists regarding the co-existence of statistics and causality to be able to provide a sound theory of scientific inference.

Acknowledgement

We are grateful to Gordon Brittan for providing helpful comments on an earlier draft.

References

Armistead, T. W.: Resurrecting the third variable: a critique of Pearl's causal analysis of Simpson's paradox. *The American Statistician*, 68(1), 1–7 (2014).

Aldrich, A.: Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10(4), 364–376 (1995).

Bandyopadhyay, P. S., Nelson, D., Greenwood, M., Brittan, G., and Berwald, J.: The logic of Simpson's paradox. *Synthese*, 181, 185–208 (2011).

Bickel, P. J., Hammel, E. A., and O'Connell, J. W.: Sex bias in graduate admissions. *Science*, 187, 398–404 (1975).

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

Chakrabarti, K. K.: *Classical Indian philosophy of induction: the Nyaya viewpoint.* Lexington Books, 2010.

Chakravarty, S. and Sarkar, P.: A paradox for inequality indices. Working Papers 559, ECINEQ, Society for the Study of Economic Inequality, 2020, `https://ideas.repec.org/p/inq/inqwps/ecineq2020-559.html`.

Cohen, M. and Nagel, E.: *An Introduction to Logic and the Scientific Method.* Harcourt, Brace and Company, New York, 1934.

Cox, D.R., and Wermuth, N.: Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*, 82(4), 688–689 (1995).

Daudt, H., and D.W. Rae, D.W.: The Ostrogorski paradox: a peculiarity of compound majority decision, *European Journal of Political Research*, 4(4), 391-399 (1976).

Fisher, R.: Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B (Methodological)*, 17(1), 69–78 (1955).

Fitelson, B.: (2017). Confirmation, causation, and Simpson's paradox. Episteme, 14(3), 297–309 (2017).

Freeman, D.: Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*, 82(4), 692–693 (1995).

Gardner, M.: "Mathematical Games." *Scientific American*, 234(3), 119–124 (1976).

Good, I. J., and Mittal. Y.: The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics*, 15(7), 694–711 (1987).

Hand, D. J.: Deconstructing statistical questions. *Journal of the Royal Statistical Society, Series A*, 157(3), 317–356 (1994).

Hawthorne, J.: Inductive Logic, The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), `https://plato.stanford.edu/archives/spr2021/entries/logic-inductive/`.

Imbens, G. W., and Rubin, D.B.: Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*, 82(4), 694–695 (1995).

Kornhauser, L. A., and Sager, L.G.: (1986). Unpacking the court, *Yale Law Journal*, 96(1), 82–117, (1986).

Levi, I.: Inductive inference as ampliative and non monotonic reasoning. *TARK'05: Proceedings of the 10th conference on Theoretical Aspects of Rationality and Knowledge*, 177–192 (2005).

Lewis, D.: Causation. *Journal of Philosophy*, 70, 556–567 (1973).

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

Linsky, B.: *Logical Constructions*. The Stanford Encyclopedia of Philosophy, `https://plato.stanford.edu/entries/logical-construction/` (2019).

List, C., and Pettit, P.:. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1), 89–110, (2002).

Mahalanobis, P.C.: Why statistics? *Sankhyā: The Indian Journal of Statistics*, 10(3), 195–228, (1950).

Mayo, D.G., and Cox, D. R.: Frequentist statistics as a theory of inductive inference, *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 49, 77–97, (2006).

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

Pagano, M., and Gauvreau, K.: *Principles of Biostatistics*, second edition. Duxbury, 1993.

Pearce, N, and Lawlor, D.: Causal inference – so much more than statistics. *International Journal of Epidemiology*, 1895–1903, (2016).

Pearl, J. and Mackenzie. D.: *The book of why*. Basic Books, New York (2018).

Pearl J, Glymour M, and Jewell N.: Causal inference in statistics: a primer. *Wiley*, UK, (2016).

Pearl, J.: Comment: understanding Simpson's paradox. *The American Statistician*, 68(1), 8–13 (2014).

Pearl, J.: *Causality: models, reasoning and inference*, second edition. Cambridge University Press, 2009.

Pearl, J.: Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688 (1995).

Pearson, K., Lee, A., and Bramley-Moore, L.: Genetic (reproductive) selection: inheritance of fertility in man, and of fecundity in thoroughbred race horses. *Philosophical Transactions of the Royal Society of London, Series A*, 192, 257–330 (1899).

Rao, C. R.: *Statistics and truth: putting chance to work*, second edition, World Scientific, 1997.

Russell, B.: On the notion of cause. In the Collected Papers of Bertrand Russell, vol 12, London: Routledge, 1992, 193–210. Originally published in 1913.

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

Spanos, A.: Yule-Simpson's paradox: the probabilistic versus the empirical conundrum. *Statistical Methods & Applications*, `https://doi.org/10.1007/s10260-020-00536-4` (2020).

Selvitella, A.: The ubiquity of the Simpson's paradox. Journal of Statistical Distributions and Applications 4, 1–16 (2017).

Selvitella, A.: The Simpson's paradox in quantum mechanics. Journal of Mathematical Physics, 58 (2017a).

Spirties, P., Glymour, C., and Scheines, R.: *Causation, prediction, and search.* MIT Press, Cambridge (2000).

Sprenger, J. and Weinberger, N.: Simpson's Paradox, The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), `https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/`.

Yule, G.: Notes on the theory of association of attributes in statistics. *Biometrika*, 2, 121–134 (1903).

A Proofs

A.1 Proof of Theorem 1

Proof. The result is proved by showing an example for each of the cases. Note that for $1 \leq i \leq 13$, Case number $(28 - i)$ transforms to Case number i by interchanging the rows of T_1, T_2 and $T_1 + T_2$. Further, by interchanging the tables T_1 and T_2, Cases 4, 5 and 6 get transformed to Cases 10, 11 and 12 respectively. In view of these observations, it is sufficient to provide examples for Cases 1 to 9 and 13, 14. Figure 2 provides these examples.

□

A.2 Proof of Theorem 2

Proof. Consider the first point.

- We show that $A_1 > C_1$, $A_2 > C_2$ and $A_1 = A_2$ together imply $\mu > \nu$ and so SP_1 cannot hold. Let $k = A_1 = A_2$. Then $a_1 = k\alpha_1$ and $a_2 = k\alpha_2$ and $\mu = (a_1 + a_2)/(\alpha_1 + \alpha_2) = k$. From $k = A_1 > C_1$, we have $k\gamma_1 > c_1$ and from $k = A_2 > C_2$, we have $k\gamma_2 > c_2$. So, $c_1 + c_2 < k(\gamma_1 + \gamma_2)$ which shows that $\mu = k > (c_1 + c_2)/(\gamma_1 + \gamma_2) = \nu$.

- A similar argument shows that $A_1 < C_1$, $A_2 < C_2$ and $A_1 = A_2$ together imply $\mu < \nu$ and so SP_2 cannot hold.

So, if $A_1 = A_2$, then neither SP_1 nor SP_2 hold and so SP does not hold. Contrapositively, if SP holds, then $A_1 \neq A_2$.

Consider the second point.

- We show that $A_1 > C_1$, $A_2 > C_2$ and $C_1 = C_2$ together imply $\mu > \nu$ and so SP_1 cannot hold. Let $s = C_1 = C_2$. Then $c_1 = s\gamma_1$ and $c_2 = s\gamma_2$ and $\nu = (c_1 + c_2)/(\gamma_1 + \gamma_2) = s$. From $A_1 > C_1 = s$, we have $a_1 > s\alpha_1$ and from $A_2 > C_2 = s$, we have $a_2 > s\alpha_2$. So, $a_1 + a_2 > s(\alpha_1 + \alpha_2)$ which shows that $\mu = (a_1 + a_2)/(\alpha_1 + \alpha_2) > s = \nu$.

Figure 2: Examples of T_1, T_2 and $T_1 + T_2$ satisfying Cases 1 to 14 of Figure 1.

Case	T_1	T_2	$T_1 + T_2$
1	$\begin{matrix} 3 & 1 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 3 & 1 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 6 & 2 \\ 2 & 2 \end{matrix}$
2	$\begin{matrix} 3 & 1 \\ 6 & 3 \end{matrix}$	$\begin{matrix} 5 & 7 \\ 1 & 4 \end{matrix}$	$\begin{matrix} 8 & 8 \\ 7 & 7 \end{matrix}$
3	$\begin{matrix} 5 & 3 \\ 10 & 10 \end{matrix}$	$\begin{matrix} 1 & 19 \\ 1 & 20 \end{matrix}$	$\begin{matrix} 6 & 22 \\ 11 & 30 \end{matrix}$
4	$\begin{matrix} 3 & 1 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 2 & 2 \\ 3 & 3 \end{matrix}$	$\begin{matrix} 5 & 3 \\ 4 & 4 \end{matrix}$
5	$\begin{matrix} 2 & 8 \\ 1 & 5 \end{matrix}$	$\begin{matrix} 4 & 4 \\ 3 & 3 \end{matrix}$	$\begin{matrix} 6 & 12 \\ 4 & 8 \end{matrix}$
6	$\begin{matrix} 2 & 1 \\ 3 & 2 \end{matrix}$	$\begin{matrix} 3 & 3 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 5 & 4 \\ 4 & 3 \end{matrix}$
7	$\begin{matrix} 10 & 5 \\ 3 & 5 \end{matrix}$	$\begin{matrix} 1 & 2 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 11 & 7 \\ 4 & 6 \end{matrix}$
8	$\begin{matrix} 2 & 1 \\ 1 & 3 \end{matrix}$	$\begin{matrix} 1 & 5 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 3 & 6 \\ 2 & 4 \end{matrix}$
9	$\begin{matrix} 3 & 2 \\ 1 & 1 \end{matrix}$	$\begin{matrix} 1 & 5 \\ 3 & 2 \end{matrix}$	$\begin{matrix} 4 & 7 \\ 4 & 3 \end{matrix}$
13	$\begin{matrix} 2 & 2 \\ 3 & 3 \end{matrix}$	$\begin{matrix} 1 & 2 \\ 2 & 4 \end{matrix}$	$\begin{matrix} 3 & 4 \\ 5 & 7 \end{matrix}$
14	$\begin{matrix} 2 & 2 \\ 3 & 3 \end{matrix}$	$\begin{matrix} 2 & 2 \\ 2 & 2 \end{matrix}$	$\begin{matrix} 4 & 4 \\ 5 & 5 \end{matrix}$

- A similar argument shows that $A_1 < C_1$, $A_2 < C_2$ and $A_1 = A_2$ together imply $\mu < \nu$ and so SP_2 cannot hold.

So, if $C_1 = C_2$, then neither SP_1 nor SP_2 hold and so SP does not hold. Contrapositively, if SP holds, then $C_1 \neq C_2$.

Consider the third point.

- We show that $A_1 > C_1$, $A_2 > C_2$, $\alpha_1 = \gamma_1$ and $\alpha_2 = \gamma_2$ together imply $\mu > \nu$ and so SP_1 cannot hold. From $A_1 > C_1$ and $\alpha_1 = \gamma_1$, we have $a_1 > c_1$; from $A_2 > C_2$ and $\alpha_2 = \gamma_2$, we have $a_2 > c_2$. So, $a_1 + a_2 > c_1 + c_2$. From $\alpha_1 = \gamma_1$ and $\alpha_2 = \gamma_2$, we have $\alpha_1 + \alpha_2 = \gamma_1 + \gamma_2 = x$ (say). Then $\mu = (a_1 + a_2)/(\alpha_1 + \alpha_2) = (a_1 + a_2)/x > (c_1 + c_2)/x = (c_1 + c_2)/(\gamma_1 + \gamma_2) = \nu$.

- A similar argument shows that $A_1 < C_1$, $A_2 < C_2$, $\alpha_1 = \gamma_1$ and $\alpha_2 = \gamma_2$ together imply $\mu < \nu$ and so SP_2 cannot hold.

So, if $\alpha_1 = \gamma_1$ and $\alpha_2 = \gamma_2$, then neither SP_1 nor SP_2 hold and so SP does not hold. Contrapositively, if SP holds, then $\alpha_1 \neq \gamma_1$ or $\alpha_2 \neq \gamma_2$. □

A.3 Proof of Theorem 3

Proof. For $i = 1, 2$, note that $\kappa(T_i) > 1$, i.e., $a_i d_i > b_i c_i$ holds if and only if

$$\frac{a_i}{a_i + b_i} \quad > \quad \frac{c_i}{c_i + d_i}. \tag{7}$$

Similarly, $\kappa(T_1 + T_2) > 1$, i.e., $(a_1 + a_2)(d_1 + d_2) > (b_1 + b_2)(c_1 + c_2)$ holds if and only if

$$\frac{a_1 + a_2}{a_1 + a_2 + b_1 + b_2} \quad > \quad \frac{c_1 + c_2}{c_1 + c_2 + d_1 + d_2}. \tag{8}$$

So, SP_1 is equivalent to the following: $(\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1) \wedge \neg (\kappa(T_1 + T_2) > 1)$. Similarly, SP_2 is equivalent to the following: $(\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1) \wedge \neg (\kappa(T_1 + T_2) < 1)$. Consequently, we have that $\mathsf{SP}_1 \vee \mathsf{SP}_2$ is equivalent to (2). □

A.4 Proof of Theorem 4

Proof. From (2), Simpson's paradox does not hold if and only if the following condition holds.

$$\neg ((\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1) \wedge \neg (\kappa(T_1 + T_2) > 1)) \text{ and } \neg ((\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1) \wedge \neg (\kappa(T_1 + T_2) < 1)).$$

Using basic logical equivalences, we obtain

$$\begin{aligned} & \neg ((\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1) \wedge \neg (\kappa(T_1 + T_2) > 1)) \\ \equiv \; & \neg (\kappa(T_1) > 1) \vee \neg (\kappa(T_2) > 1) \vee (\kappa(T_1 + T_2) > 1) \\ \equiv \; & \neg ((\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1)) \vee (\kappa(T_1 + T_2) > 1) \\ \equiv \; & ((\kappa(T_1) > 1) \wedge (\kappa(T_2) > 1)) \text{ implies } (\kappa(T_1 + T_2) > 1). \end{aligned}$$

Similarly, one can argue that the condition $\neg ((\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1) \wedge \neg (\kappa(T_1 + T_2) < 1))$ is equivalent to $(\kappa(T_1) < 1) \wedge (\kappa(T_2) < 1)$ implies $\kappa(T_1 + T_2) < 1$. □

A.5 Proof of Theorem 5

Proof. We need to show that if any one of the conditions given in Definition 4 hold, then SP does not occur. We show this for the first condition. The proofs for the other conditions are similar.

Consider the first condition, namely, $\max(a_1/b_1, a_2/b_2) < \min(c_1/d_1, c_2/d_2)$. This is equivalent to the following four conditions: $a_1d_1 < b_1c_1$, $a_2d_2 < b_2c_2$, $a_1d_2 < c_2b_1$ and $a_2d_1 < c_1b_2$. From $a_1d_1 < b_1c_1$ and $a_2d_2 < b_2c_2$ we have $\kappa(T_1) < 1$ and $\kappa(T_2) < 1$. So, the antecedent of (3) is false and so (3) is true. The antecedent of (4), on the other hand, is true. The consequent of (4) is $\kappa(T_1 + T_2) < 1$. The last condition is equivalent to $(a_1 + a_2)(d_1 + d_2) < (b_1 + b_2)(c_1 + c_2)$. From the four inequalities indentified above, it follows that each of the cross product term in $(a_1 + a_2)(d_1 + d_2)$ is less than a corresponding and unique cross-product term in $(b_1 + b_2)(c_1 + c_2)$. So, we have $\kappa(T_1 + T_2) < 1$ and therefore (4) is true. Consequently, SP does not hold. □

A.6 Proof of Theorem 6

Proof. The WORH states that either $\kappa(T_1) = \kappa(T_1 + T_2)$ or $\kappa(T_2) = \kappa(T_1 + T_2)$. Suppose that $\kappa(T_1) = \kappa(T_1 + T_2)$, the other case being similar.

If either $\kappa(T_1) = 1$ or $\kappa(T_2) = 1$, then the antecedents of both (3) and (4) are false and consequently, both (3) and (4) are true. So, suppose that $\kappa(T_1) \neq 1$ and $\kappa(T_2) \neq 1$. This leads to four cases, namely, Case-1: ($\kappa(T_1) > 1$ and $\kappa(T_2) > 1$); Case-2: ($\kappa(T_1) > 1$ and $\kappa(T_2) < 1$); Case-3: ($\kappa(T_1) < 1$ and $\kappa(T_2) > 1$); Case-4: ($\kappa(T_1) < 1$ and $\kappa(T_2) < 1$). For Cases-2 and 3, the antecedents of both (3) and (4) are false and consequently, both (3) and (4) are true. Now Consider Case-1. The antecedent of (4) is false and so (4) is true. On the other hand, the antecedent of (3) is true. Using $\kappa(T_1) = \kappa(T_1 + T_2)$, we have $\kappa(T_1 + T_2) > 1$ so that the consequent of (3) is also true. So, (3) is true. So, in Case-1, both (3) and (4) are true and therefore SP does not hold. The argument for Case-4 is similar. □

A.7 Proof of Theorem 7

Proof. The random variables X and Y are positively associated if $\Pr[Y|X] > \Pr[Y|\overline{X}]$, i.e., if $a/(a+b) > c/(c + d)$. The last condition holds if and only if $ad \geq bc$, i.e., if and only if $\kappa(T) > 1$. The argument for the positive association of X and $\overline{Y}$ is similar. □

A.8 Proof of Theorem 9

Proof. To start with we translate the given conditions in terms of a_i, b_i, c_i and d_i, $i = 1, 2$. The conditions $(X \sim Y)|M$ and $(X \sim Y)|\overline{M}$ both hold, so using Theorem 7, we have $\kappa(T_i) > 1$, $i = 1, 2$. It is given that SP holds for (T_1, T_2). From Theorem 3, it follows that $\kappa(T_1 + T_2) \leq 1$. Note that $\kappa(T_i) > 1$ is equivalent to the condition $a_id_i > b_ic_i$ and $\kappa(T_1 + T_2) \leq 1$ is equivalent to the condition $(a_1 + a_2)(d_1 + d_2) \leq (b_1 + b_2)(c_1 + c_2)$. It is possible to write these conditions in two different ways.

Case $a_i/c_i > b_i/d_i$, $i = 1, 2$ and $(a_1+a_2)/(c_1+c_2) \leq (b_1+b_2)/(d_1+d_2)$: One may write $(a_1+a_2)/(c_1+c_2) = \lambda(a_1/c_1)+(1-\lambda)(a_2/c_2)$ where $\lambda = c_1/(c_1+c_2)$ and $(b_1+b_2)/(d_1+d_2) = \mu(b_1/d_1)+(1-\mu)(b_2/d_2)$ where $\mu = d_1/(d_1 + d_2)$. So, in particular $(a_1 + a_2)/(c_1 + c_2)$ lies in the interval defined by a_1/c_1 and a_2/c_2 and $(b_1 + b_2)/(d_1 + d_2)$ lies in the interval defined by b_1/d_1 and b_2/d_2.

Since $a_i/c_i > b_i/d_i$, $i = 1, 2$, and $(a_1 + a_2)/(c_1 + c_2) \leq (b_1 + b_2)/(d_1 + d_2)$, we have $\mu(b_1/d_1) + (1 - \mu)(b_2/d_2) = (b_1+b_2)/(d_1+d_2) \geq (a_1+a_2)/(c_1+c_2) = \lambda(a_1/c_1)+(1-\lambda)(a_2/c_2) > \lambda(b_1/d_1)+(1-\lambda)(b_2/d_2)$. So, $(\mu - \lambda)(b_1/d_1 - b_2/d_2) > 0$ and consequently, $\mu > \lambda$ or $\mu < \lambda$ according as $b_1/d_1 > b_2/d_2$ or

$b_1/d_1 < b_2/d_2$; in particular, neither $\mu = \lambda$ nor $b_1/d_1 = b_2/d_2$ are possible. Using the definitions of μ and λ, we have $c_1/c_2 < d_1/d_2$ or $c_1/c_2 > d_1/d_2$ according as $b_1/d_1 > b_2/d_2$ or $b_1/d_1 < b_2/d_2$.

The condition $(a_1 + a_2)/(c_1 + c_2) \leq (b_1 + b_2)/(d_1 + d_2)$ forces the intervals $I = [b_1/d_1, a_1/c_1]$ and $J = [b_2/d_2, a_2/c_2]$ to be disjoint. To see this, first suppose that $a_2/c_2 \geq a_1/c_1$. If the intervals I and J are not disjoint, then we have the relation $b_1/d_1 \leq b_2/d_2 \leq a_1/c_1 \leq a_2/c_2$. Since $(a_1 + a_2)/(c_1 + c_2)$ lies in $[a_1/c_1, a_2/c_2]$ and $(b_1 + b_2)/(d_1 + d_2)$ lies in $[b_1/d_1, b_2/d_2]$, it follows that $(a_1 + a_2)/(c_1 + c_2) \geq (b_1+b_2)/(d_1+d_2)$. Equality occurs if and only if $(b_1+b_2)/(d_1+d_2) = b_2/d_2 = a_1/c_1 = (a_1+a_2)/(c_1+c_2)$, i.e., if and only if $b_1/d_1 = b_2/d_2 = a_1/c_1 = a_2/c_2$. This in particular means that $b_1/d_1 = a_1/c_1$. Since it is given that $a_1/c_1 > b_1/d_1$, equality cannot occur and so, $(a_1 + a_2)/(c_1 + c_2) > (b_1 + b_2)/(d_1 + d_2)$. The last condition contradicts $(a_1 + a_2)/(c_1 + c_2) \leq (b_1 + b_2)/(d_1 + d_2)$. Similarly, if $a_1/c_1 \geq a_2/c_2$, then also we obtain a contradiction. This establishes that I and J are disjoint intervals. So, there are two possibilities: (A) $b_1/d_1 < a_1/c_1 < b_2/d_2 < a_2/c_2$, or (B) $b_2/d_2 < a_2/c_2 < b_1/d_1 < a_1/c_1$. From the previous discussion, we obtain that in (A), $c_1/c_2 > d_1/d_2$ holds, while in (B) $c_1/c_2 < d_1/d_2$ holds.

Case $a_i/b_i > c_i/d_i$, $i = 1, 2$ and $(a_1 + a_2)/(b_1 + b_2) \leq (c_1 + c_2)/(d_1 + d_2)$: As argued above, note that $(a_1 + a_2)/(b_1 + b_2)$ lies in the interval defined by a_1/b_1 and a_2/b_2 and similarly, $(c_1 + c_2)/(d_1 + d_2)$ lies in the interval defined by c_1/d_1 and c_2/d_2. Also, arguing as above, the condition $(a_1 + a_2)/(b_1 + b_2) \leq (c_1 + c_2)/(d_1 + d_2)$ forces the intervals $[c_1/d_1, a_1/b_1]$ and $[c_2/d_2, a_2/b_2]$ to be disjoint. This leads to the following two possibilities: (C) $c_1/d_1 < a_1/b_1 < c_2/d_2 < a_2/b_2$, or (D) $c_2/d_2 < a_2/b_2 < c_1/d_1 < a_1/b_1$. In (C), we have $c_1/c_2 < d_1/d_2$, while in (D) we have $c_1/c_2 > d_1/d_2$.

The combined effect of the above two cases is that if $\kappa(T_i) > 1$, $i = 1, 2$ and $\kappa(T_1 + T_2) \leq 1$, then either (A) and (C) both occur, or (B) and (D) both occur. Since $(a_2+b_2)/(c_2+d_2)$ lies in $[b_2/d_2, a_2/c_2]$ and $(a_1 + b_1)/(c_1 + d_1)$ lies in $[b_1/d_1, a_1/c_1]$, (B) implies $(a_2 + b_2)/(c_2 + d_2) < (a_1 + b_1)/(c_1 + d_1)$ which is equivalent to $M \sim X$. A similar reasoning shows that (D) implies $M \sim Y$. So, if (B) and (D) hold, then both $M \sim X$ and $M \sim Y$ hold. On the other hand, an analogous reasoning shows that if (A) and (C) hold, then both $M \sim \overline{X}$ and $M \sim \overline{Y}$ hold. □

A.9 Proof of Theorem 10

Proof. For a non-zero δ and a 2×2 table T, by δT, we denote the 2×2 table obtained from T by multiplying each entry with δ. It is easy to see that for 2×2 tables T_1 and T_2, if (T_1, T_2) is an instance of positive SP, then so is $(\delta T_1, \delta T_2)$. Similarly, for positive alignment, negative SP and negative alignment.

Define

$$T_1^{(0)} = \begin{array}{|c|c|}\hline 5 & 3 \\ \hline 10 & 10 \\ \hline\end{array}, \quad T_2^{(0)} = \begin{array}{|c|c|}\hline 1 & 19 \\ \hline 1 & 20 \\ \hline\end{array}, \quad T_1^{(0)} + T_2^{(0)} = \begin{array}{|c|c|}\hline 6 & 22 \\ \hline 11 & 30 \\ \hline\end{array}.$$

$$T_1^{(1)} = \begin{array}{|c|c|}\hline 10 & 6 \\ \hline 10 & 10 \\ \hline\end{array}, \quad T_2^{(1)} = \begin{array}{|c|c|}\hline 1 & 19 \\ \hline 1 & 20 \\ \hline\end{array}, \quad T_1^{(1)} + T_2^{(1)} = \begin{array}{|c|c|}\hline 11 & 25 \\ \hline 11 & 30 \\ \hline\end{array}.$$

$$T_1^{(2)} = \begin{array}{|c|c|}\hline 100 & 60 \\ \hline 20 & 10 \\ \hline\end{array}, \quad T_2^{(2)} = \begin{array}{|c|c|}\hline 1 & 19 \\ \hline 5 & 20 \\ \hline\end{array}, \quad T_1^{(2)} + T_2^{(2)} = \begin{array}{|c|c|}\hline 101 & 79 \\ \hline 25 & 30 \\ \hline\end{array}.$$

$$T_1^{(3)} = \begin{array}{|c|c|}\hline 100 & 60 \\ \hline 20 & 10 \\ \hline\end{array}, \quad T_2^{(3)} = \begin{array}{|c|c|}\hline 20 & 90 \\ \hline 5 & 20 \\ \hline\end{array}, \quad T_1^{(3)} + T_2^{(3)} = \begin{array}{|c|c|}\hline 120 & 150 \\ \hline 25 & 30 \\ \hline\end{array}.$$

Note that $(T_1^{(0)}, T_2^{(0)})$ is an instance of positive SP, $(T_1^{(1)}, T_2^{(1)})$ is an instance of positive alignment,

$(T_1^{(2)}, T_2^{(2)})$ is an instance of negative SP, and $(T_1^{(3)}, T_2^{(3)})$ is an instance of negative aligment.

For $k \geq 4$, define $(T_1^{(k)}, T_2^{(k)}) = (\delta T_1^{(k-4)}, \delta T_2^{(k-4)})$, where $\delta = 20$. Then $\mathcal{T} = (T_1^{(k)}, T_2^{(k)})_{k \geq 0}$ is a monotonic sequence possessing the properties mentioned in the theorem. $\square$