Understanding Simpson's Paradox

Judea Pearl
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095-1596
judea@cs.ucla.edu
(310) 825-3243 Tel / (310) 794-5057 Fax

Simpson's paradox is often presented as a compelling demonstration of why we need statistics education in our schools. It is a reminder of how easy it is to fall into a web of paradoxical conclusions when relying solely on intuition, unaided by rigorous statistical methods. In recent years, ironically, the paradox assumed an added dimension when educators began using it to demonstrate the limits of statistical methods, and why causal, rather than statistical considerations are necessary to avoid those paradoxical conclusions (Arah, 2008; Pearl, 2009, pp. 173–182; Wasserman, 2004).

My comments are divided into two parts. First, I will give a brief summary of the history of Simpson's paradox and how it has been treated in the statistical literature in the past century. Next I will ask what is required to declare the paradox "resolved," and argue that modern understanding of causal inference has met those requirements.

1 The History

Simpson's paradox refers to a phenomena whereby the association between a pair of variables (X, Y) reverses sign upon conditioning of a third variable, Z, regardless of the value taken by Z. If we partition the data into subpopulations, each representing a specific value of the third variable, the phenomena appears as a sign reversal between the associations measured in the disaggregated subpopulations relative to the aggregated data, which describes the population as a whole.

Edward H. Simpson first addressed this reversal in a technical paper in 1951 (Simpson, 1951), but Karl Pearson et al. (1899) and Udny Yule (1903) had mentioned similar effects half a century earlier, reporting disappearing associations, rather than reversing signs.

Chapter 6 of my book *Causality* (Pearl, 2009, p. 176) remarks that, surprisingly, only two articles in the statistical literature attribute the peculiarity of Simpson's reversal to causal interpretations. The first is Pearson et al. (1899), in which a short remark warns us that correlation is not causation, and the second is Lindley and Novick (1981) who mentioned the possibility of explaining the paradox in "the language of causation" but chose not to do so "because the concept, although widely used, does not seem to be well defined" (p. 51).

My survey further documents that, other than these two exceptions, the entire statistical literature from Pearson et al. (1899) to the 1990s was not prepared to accept the idea that a statistical peculiarity, so clearly demonstrated in the data, could have causal roots.[1]

In particular, the word "causal" does not appear in Simpson's paper, nor in the vast literature that followed, including Blyth (1972), who coined the term "paradox," and the influential writings of Agresti (1983), Bishop et al. (1975), and Whittemore (1978).

What Simpson did notice though, was that depending on the story behind the data, the more "sensible interpretation" (his words) is sometimes compatible with the aggregate population, and sometimes with the disaggregated subpopulations. His example of the former involved a deck of cards, in which two face types become associated when partitioned according to a cleverly crafted rule. This key observation remained unnoticed until Lindley and Novick (1981) replicated it using a more realistic example. The idea that statistical data, however large, is insufficient for determining what is "sensible," and that it must be supplemented with extra-statistical knowledge to make sense was considered heresy in the 1950s.

Lindley and Novick (1981) elevated Simpson's paradox to new heights by showing that there was no statistical criterion that would warn the investigator against drawing the wrong conclusions or indicate which data represented the correct answer. First they showed that reversal may lead to difficult choices in critical decision-making situations:

> "The apparent answer is, that when we know that the gender of the patient is male or when we know that it is female we do not use the treatment, but if the gender is unknown we should use the treatment! Obviously that conclusion is ridiculous." (Novick, 1983, p. 45)

Second, they showed that, with the very same data, we should consult either the combined table or the disaggregated tables, depending on the context. Clearly, when two different contexts compel us to take two opposite actions based on the same data, our decision must be driven not by statistical considerations, but by some additional information extracted from the context.

Thirdly, they postulated a scientific characterization of the extra-statistical information that researchers take from the context, and which causes them to form a consensus as to which table gives the correct answer. That Lindley and Novick opted to characterize this information in terms of "exchangeability" rather than causality is understandable;[2] the state of causal language in the 1980s was so primitive that they could not express even the simple yet crucial fact that gender is not affected by the treatment.[3] What is important though, is that the example they used to demonstrate that the correct answer lies in the aggregated

[1] This stands in sharp contrast to the claims of Hernán et al. (2011) according to which causal explanations of Simpson's paradox had been "formally described and explained in causal terms half a century before the publication of Simpson's article." Simpson and his predecessor did not have the language to articulate, let alone explain causal phenomena.

[2] Lindley later regretted that choice (Pearl, 2009, p. 384), and indeed, his treatment of exchangeability was guided exclusively by causal considerations (Meek and Glymour, 1994).

[3] Statistics teachers would enjoy the challenge of explaining how the sentence "treatment does not change gender" can be expressed mathematically. Lindley and Novick tried, unsuccessfully of course, to use conditional probabilities.

data, had a totally different causal structure than the one where the correct answer lies in the disaggregated data. Specifically, the third variable (Plant Height) was affected by the treatment (Plant Color) as opposed to Gender which is a pre-treatment confounder. (See an isomorphic model in Fig. 1(b), where Blood-pressure replacing Plant-Height.)

More than 30 years have passed since the publication of Lindley and Novick's paper, and the face of causality has changed dramatically. Not only do we now know which causal structures would support Simpson's reversals, we also know which structure places the correct answer with the aggregated data or with the disaggregated data. Moreover, the criterion for predicting where the correct answer lies (and, accordingly, where human consensus resides) turns out to be oblivious to temporal information, nor does it depend on whether or not the third variable is affected by the treatment. It involves a simple graphical condition called "back-door" (Pearl, 1993) which traces paths in the causal diagram and assures that all spurious paths from treatment to outcome are intercepted by the third variable. This will be demonstrated in the next section, where we argue that, armed with these criteria, we can safely proclaim Simpson's paradox "resolved."

2 A Paradox Resolved

Any claim to a resolution of a paradox, especially one that has resisted a century of attempted resolution must meet certain criteria. First and foremost, the solution must explain why people consider the phenomenon surprising or unbelievable. Second, the solution must identify the class of scenarios in which the paradox may surface, and distinguish it from scenarios where it will surely not surface. Finally, in those scenarios where the paradox leads to indecision, we must identify the correct answer, explain the features of the scenario that lead to that choice, and prove mathematically that the answer chosen is indeed correct. The next three subsections will describe how these three requirements are met in the case of Simpson's paradox and, naturally, will proceed to convince readers that the paradox deserves the title "resolved."

2.1 Simpson's Surprise

In explaining the surprise, we must first distinguish between "Simpson's reversal" and "Simpson's paradox"; the former being an arithmetic phenomenon in the calculus of proportions, the latter a psychological phenomenon that evokes surprise and disbelief. A full understanding of Simpson's paradox should explain why an innocent arithmetic reversal of an association, albeit uncommon, came to be regarded as "paradoxical," and why it has captured the fascination of statisticians, mathematicians and philosophers for over a century (though it was first labeled "paradox" by Blyth (1972)).

The arithmetics of proportions has its share of peculiarities, no doubt, but these tend to become objects of curiosity once they have been demonstrated and explained away by examples. For instance, naive students of probability may expect the average of a product to equal the product of the averages but quickly learn to guard against such expectations, given a few counterexamples. Likewise, students expect an association measured in a mixture distribution to equal a weighted average of the individual associations. They are surprised,

therefore, when ratios of sums, $(a+b)/(c+d)$, are found to be ordered differently than individual ratios, a/c and b/d.[4] Again, such arithmetic peculiarities are quickly accommodated by seasoned students as reminders against simplistic reasoning.

In contrast, an arithmetic peculiarity becomes "paradoxical" when it clashes with deeply held convictions that the pecularity is impossible, and this occurs when one takes seriously the causal implications of Simpson's reversal in decision-making contexts. Reversals are indeed impossible whenever the third variable, say age or gender, stands for a pre-treatment covariate because, so the reasoning goes, no drug can be harmful to both males and females yet beneficial to the population as a whole. The universality of this intuition reflects a deeply held and valid conviction that such a drug is physically impossible. Remarkably, such impossibility can be derived mathematically in the calculus of causation in the form of a "sure-thing" theorem (Pearl, 2009, p. 181):

> "An action A that increases the probability of an event B in each subpopulation (of C) must also increase the probability of B in the population as a whole, provided that the action does not change the distribution of the subpopulations."[5]

Thus, regardless of whether effect size is measured by the odds ratio or other comparisons, regardless of whether Z is a confounder or not, and regardless of whether we have the correct causal structure on hand, our intuition should be offended by any effect reversal that appears to accompany the aggregation of data.

I am not aware of another condition that rules out effect reversal with comparable assertiveness and generality, requiring only that Z not be affected by our action, a requirement satisfied by all treatment-independent covariates Z. Thus, 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.

2.2 Which scenarios invite reversals?

Attending to the second requirement, we need first to agree on a language that describes and identifies the class of scenarios for which association reversal is possible. Since the notion of "scenario" connotes a process by which data is generated, a suitable language for such a process is a causal diagram, as it can simulate any data-generating process that operates sequentially along its arrows. For example, the diagram in Fig. 1(a) can be regarded as a blueprint for a process in which $Z = \mathit{Gender}$ receives a random value (male or female) depending on the gender distribution in the population. The treatment is then assigned a value (treated or untreated) according to the conditional distribution $P(\mathit{treatment}|\mathit{male})$ or $P(\mathit{treatment}|\mathit{female})$. Finally, once Gender and Treatment receive their values, the outcome process (Recovery) is activated, and assigns a value to Y using the conditional distribution $P(Y = y|X = x, Z = z)$. All these local distributions can be estimated from the data. Thus,

[4] In Simpson's paradox we witness the simultaneous orderings: $(a1+b1)/(c1+d1) > (a2+b2)/(c2+d2)$, $(a1/c1) < (a2/c2)$, and $(b1/d1) < (b2/d2)$.

[5] The no-change provision is probabilistic; it permits the action to change the classification of individual units so long as the relative sizes of the subpopulations remain unaltered.

the scientific content of a given scenario can be encoded in the form of a directed acyclic graph (DAG), capable of simulating a set of data-generating processes compatible with the given scenario.

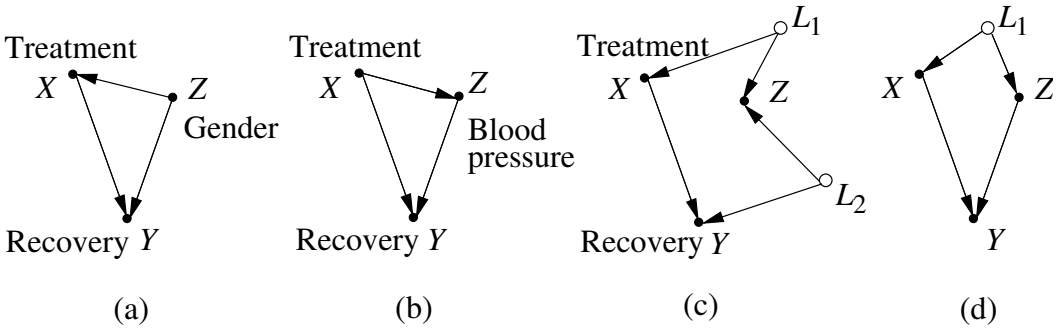

Figure 1: Graphs demonstrating the insufficiency of chronological information. In models (c) and (d), Z may occur before or after the treatment, yet the correct answer remains the same: We should not condition in (c), but condition in (d). (Hollow circles represent unmeasured latent variables.)

The theory of graphical models (Pearl, 1988; Lauritzen, 1996) can tell us, for every given DAG, whether Simpson's reversal is realizable or logical impossible in the simulated scenario. By a logical impossibility we mean that for every scenario that fits the DAG structure, there is no way to assign processes to the arrows and generate data that exhibit association reversal as described by Simpson.

For example, the theory immediately tells us that all structures depicted in Fig. 1 can exhibit reversal, while in Fig. 2, reversal can occur in (a), (b), and (c), but not in (d), (e), or (f). That Simpson's paradox can occur in each of the structures in Fig. 1 follows from the fact that the structures are observationally equivalent; each can emulate any distribution generated by the others. Therefore, if association reversal is realizable in one of the structures, say (a), it must be realizable in all structures. The same consideration applies to graphs (a), (b), and (c) of Fig. 2, but not to (d), (e), or (f) which are where the X, Y association is collapsible over Z.

2.3 Making the correct decision

We now come to the hardest test of having resolved the paradox: proving that we can make the correct decision when reversal occurs. This can be accomplished either mathematically or by simulation. Mathematically, we use an algebraic method called "*do*-calculus" (Pearl, 2009, p. 85–89) which is capable of determining, for any given model structure, the causal effect of one variable on another and which variables need to be measured to make this determination. Compliance with do-calculus should then constitute a proof that the decisions we made using graphical criteria is correct. Since some readers of this article may not be familiar with the *do*-calculus, simulation methods may be more convincing.??Simulation "proofs" should be organized as a successful "guessing game." Given a scenario, we construct a DAG that captures that scenario and use a graphical test (called "back-door") to guess

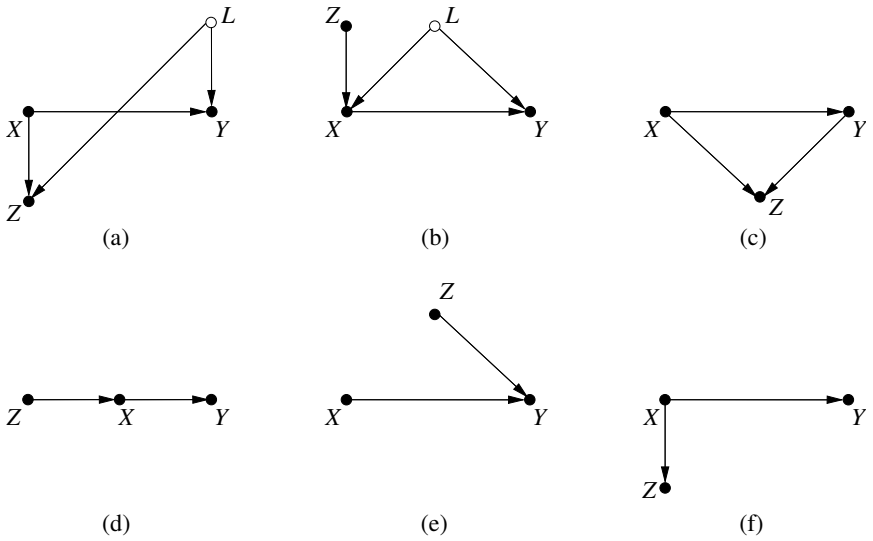

Figure 2: Simpson reversal can be realized in models (a), (b), and (c) but not in (d), (e), or (f).

where the correct answer lies. We now test whether our guess was correct by evaluating the treatment efficacy by simulating a randomized experiment on the population generated by the DAG. For example, the back-door criterion instructs us to guess that in Fig. 1, in models (b) and (c) the correct answer is provided by the aggregated data, while in structures (a) and (d) the correct answer is provided by the disaggregated data. We simulate a randomized experiment on the (fictitious) population to determine whether the resulting effect is positive or negative, and compare it with the associations measured in the aggregated and disaggregated population.??Remarkably, our guesses should prove correct regardless of the parameters used in the simulation model, as along as the structure is obeyed. This explains how people form a consensus about which data is "more sensible (Simpson, 1951) prior to actually seeing the data.

This is a good place to explain how the back-door criterion works, and how it determines where the correct answer resides. The principle is simple: The paths connecting X and Y are of two kinds, causal and spurious. Causative associations are carried by the causal paths, namely, those tracing arrows directed from X to Y. The other paths carry spurious associations and need to be blocked by conditioning on an appropriate set of covariates. All paths containing an arrow into X are spurious paths, and need to be intercepted by the chosen set of covariates.

When dealing with a singleton covariate Z, as in the Simpson's paradox, we need to merely ensure that

1. Z is not a descendant of X, and

2. Z blocks every path that ends with an arrow into X.

(Extensions for descendants of X are given in (Pearl, 2009, p. 338; Shpitser et al., 2010).)

The operation of "blocking" requires a special handling of "collider" variables, which behave oppositely to arrow-emitting variables. The latter block the path when conditioned on, while the former block the path when they and all their descendants are not conditioned on. This special handling of " colliders," reflects a general phenomenon known as Berkson's paradox (Berkson, 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

Armed with this criterion we can determine, for example, that in Fig. 1(b) and (d), if we wish to correctly estimate the effect of X on Y, we need to condition on Z (thus blocking the back-door path $X \leftarrow Z \rightarrow Y$). We can similarly determine that we should not condition on Z in Fig. 1(a) and (c). The former because there are no back-door paths requiring blockage, and the latter because the back-door path $X \leftarrow \circ \rightarrow Z \leftarrow \circ \rightarrow Y$ is blocked when Z is not conditioned on. The correct decisions follow from this determination; when conditioning on Z is required, the Z-specific data carries the correct information, and when conditioning on Z should be avoided, the aggregated data carries the correct information.

Finally, we should remark that, in certain models the correct answer may not lie in either the disaggregated or the aggregated data. This occurs when Z is not sufficient to block all back-door paths as in Fig. 2(b) and (c); in such cases a set of additional covariates may be needed, which takes us beyond the scope of this note.

The model in Fig. 3 presents opportunities to simulate successive reversals, which could serve as an effective (and fascinating) instruction tool for introductory statistics classes. Here we see that to block the only unblocked back-door path $X \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$, we need to condition on Z_1. This means that, if the simulation machine is set to generate association reversal, the correct answer will reside in the disaggregated, Z_1-specific data. If we further condition on a second variable, Z_2, the back-door path $X \leftarrow \circ \rightarrow Z_2 \leftarrow Z_3 \rightarrow Y$ will become unblocked, and a bias will be created, meaning that the correct answer lies with the aggregated data. Upon further conditioning on Z_3 the bias is removed and the correct answer returns to the disaggregated, Z_3-specific data.

Note that in each stage, we can set the numbers in the simulation machine so as to generate association reversal between the pre-conditioning and post-conditioning data. Note further that at any stage of the process we can check where the correct answer lies by subjecting the population generated to a hypothetical randomized trial.

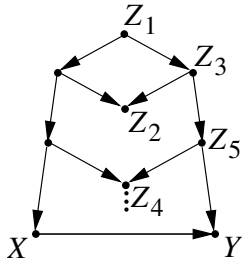

Figure 3: A multi-stage Simpson's paradox machine. Successive conditioning in the order $(Z_1, Z_2, Z_3, Z_4, Z_5)$ creates reversal at each stage, with the correct answers alternating between disaggregated and aggregated data.

3 Conclusions

I hope that playing the multi-stage Simpson's guessing game (Fig. 3) would convince readers that we now understand most of the intricacies of Simpson's paradox.

References

Agresti, A. (1983). Fallacies, statistical. In *Encyclopedia of Statistical Science* (S. Kotz and N. Johnson, eds.), vol. 3. John Wiley, New York, 24–28.

Arah, O. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology* **4** doi:10.1186/1742–7622–5–5. Online at <http://www.ete-online.com/content/5/1/5>.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2** 47–53.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, MA.

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

Hernán, M., Clayton, D. and Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology* DOI:10.1093/ije/dyr041.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford. Reprinted 2004 with corrections.

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

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

Novick, M. (1983). The centrality of Lord's paradox and exchangeability for all statistical inference. In *Principals of modern psychological measurement* (H. Wainer and S. Messick, eds.). Earlbaum, Hillsdale, NJ, 41–53.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearson, K., Lee, A. and Bramley-Moore, L. (1899). Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Transactions of the Royal Society A* **73** 534–539.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

Shpitser, I., VanderWeele, T. and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 527–536.

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

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science+Business Media, Inc., New York, NY.

Whittemore, A. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, B* **40** 328–340.

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