Comment: Understanding Simpson's Paradox

Judea PEARL

I thank the editor, Ronald Christensen, for the opportunity to discuss this important topic and to comment on the article by Armistead. 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 (Wasserman 2004; Arah 2008; Pearl 2009, pp. 173–182).

My comments are divided into three 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. Finally, I will answer specific questions raised in Armistead's article and show how the resolution of Simpson's paradox can be taught for fun and progress.

1. THE HISTORY

Simpson's paradox refers to a phenomenon 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 phenomenon 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 phenomenon in a technical article in 1951, but Karl Pearson et al. in 1899 and Udny Yule in 1903 had mentioned a similar effect earlier. All three reported associations that disappear, rather than reversing signs upon aggregation. Sign reversal was first noted by Cohen and Nagel (1934) and then by Blyth (1972) who labeled the reversal "paradox," presumably because the surprise that association reversal evokes among the unwary appears paradoxical at first.

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, Lee, and Bramley-Moore (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, Lee, and Bramley-Moore (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 article, nor in the vast literature that followed, including Blyth (1972), who coined the term "paradox," and the influential writings of Agresti (1983), Bishop, Fienberg, and Holland (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 latter involves a positive association between treatment and survival both among males and females which disappears in the combined population. Here, his "sensible interpretation" is unambiguous: "The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females." His example of the former involved a deck of cards, in which two independent face types become associated when partitioned according to a cleverly crafted rule (see Hernán, Clayton, and Keiding 2011). Here, claims Simpson, "it is the combined table which provides what we would call the sensible answer." This key observation remained unnoticed until Lindley and Novick (1981) replicated it in a more realistic example which gave rise to reversal. The idea that statistical data, however large, are 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.

Judea Pearl, Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095-1596 (E-mail: *judea@cs.ucla.edu*).This research was supported in parts by grants from NSF #IIS1249822 and #IIS1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933. I appreciate the encouragement of Ronald Christensen, conversations with Miguel Hernán, and editorial comments by Madlyn Glymour.

[1] This contrasts the historical account of Hernán, Clayton, and Keiding (2011) according to which "Such discrepancy [between marginal and conditional associations in the presence of confounding] had been already noted, 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 vocabulary to articulate, let alone formally describe and explain causal phenomena.

  DOI: 10.1080/00031305.2014.876829

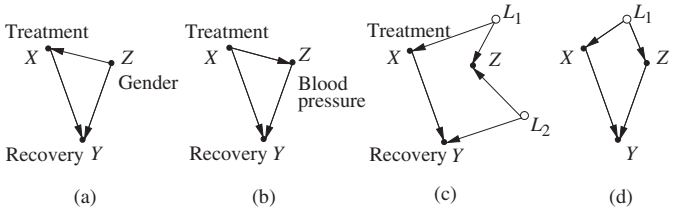

Figure 1. Graphs demonstrating the insufficiency of chronological information. In models (c) and (d), Z may occur before or after the treatment, yet the correct answer remains invariant to this timing: we should not condition on Z in model (c), and we should condition on Z in model (d). In both models, Z is not affected by the treatment.

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.

Third, 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 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 pretreatment confounder. (See an isomorphic model in Figure 1(b), where blood-pressure replacing plant-height.[4])

More than 30 years have passed since the publication of Lindley and Novick's article, 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 rather insensitive to temporal information, nor does it hinge critically 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.[5] Again, such arithmetic

[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.

[4] Interestingly, Simpson's examples also had different causal structure; in the former, the third variable (gender) was a common cause of the other two, whereas in the latter, the third variable (paint on card) was a common effect of the other two (Hernán, Clayton, and Keiding 2011). Yet, although this difference changed Simpson's intuition of what is "more sensible," it did not stimulate his curiosity as a fundamental difference, worthy of scientific exploration.

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

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 peculiarity 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 pretreatment 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.[6]

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 Figure 1(a) can be regarded as a blueprint for a process in which $Z =$ 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 $P(\text{treatment}|\text{male})$ or $P(\text{treatment} \mid \text{female})$. 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, 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.

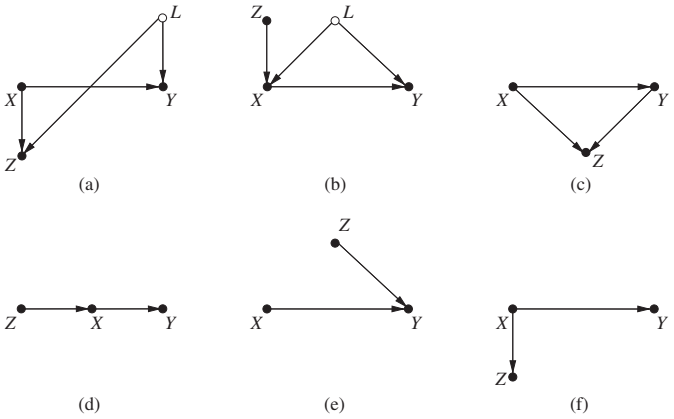

Figure 2. Simpson's reversal can be realized in models (a), (b), and (c) but not in (d), (e), or (f).

The theory of graphical models (Pearl 1988; Lauritzen 1996) can tell us, for a given DAG, whether Simpson's reversal is realizable or logically 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 Figure 1 can exhibit reversal, while in Figure 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 Figure 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 Figure 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.[7] 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" can be organized as a "guessing game," where a "challenger" who knows the model behind the data dares an analyst to guess what the causal effect is (of X on Y) and checks the answer against

[6] 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.

[7] When such determination cannot be made from the given graph, as is the case in Figure 2(b), the do-calculus alerts us to this fact.

the gold standard of a randomized trial, simulated on the model. Specifically, the "challenger" chooses a scenario (or a "story" to be simulated), and a set of simulation parameters such that the data generated would exhibit Simpson's reversal. He then reveals the scenario (not the parameters) to the analyst. The analyst constructs a DAG that captures the scenario and guesses (using the structure of the DAG), whether the correct answer lies in the aggregated or disaggregated data. Finally, the "challenger" simulates a randomized trial on a fictitious population generated by the model, estimates the underlying causal effect, and checks the result against the analyst's guess.

For example, the back-door criterion instructs us to guess that in Figure 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 long as the structure of the simulator remains the same.[8] 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, VanderWeele, and Robins (2010), and Pearl and Paz (2013)).

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 Figures 1(a) 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 Figures 1(b) 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 carry the correct information. In Figure 2(c), for example, the aggregated information carries the correct information because the spurious (noncausal) path $X \rightarrow Z \leftarrow Y$ is blocked when Z is not conditioned on. The same applies to Figures 2(a) and 1(c).

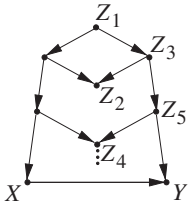

Figure 3. A multistage Simpson's paradox machine. Commulative conditioning in the order (Z_1, Z_2, Z_3, Z_4, Z_5) creates reversal at each stage, with the correct answers alternating between disaggregated and aggregated data.

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 an active back-door path as in Figure 2(b); in such cases, a set of additional covariates may be needed, which takes us beyond the scope of this note.

The model in Figure 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 preconditioning 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.

3. ARMISTEAD'S CRITIQUE

Armistead does not disagree with the technical points presented above and rightly so; they are backed by sound mathematical proofs. The main point of contention seems to be whether the disaggregated data are still valuable, when the correct answer lies with the aggregated data (as in Figures 1(a) and (c)). On this issue, Armistead says:

[8] By "structure" we mean the list of variables that need be consulted in computing each variable V_i in the simulation.

> Whether causal or not, third variables can convey critical information about a first order relationship, study design, and previously unobserved variables. Any conditioning on nontrivial third variable that produces Simpson's Paradox should be carefully examined before either the aggregated or the disaggregated findings are accepted, regardless of whether the variable is thought to be causal.

I agree with the general thrust of this paragraph. Every variable can indeed "convey critical information" if such information is needed for answering the investigator's research question. But in our examples, we asked not whether the third variable conveys information about study design or other interesting subjects; we asked whether it would help us estimate the total effect of X on Y. In the context of this query, the answer is: NO; the aggregated (or disaggregated) findings can be accepted without further examination.

When we endeavor to ask other queries, other than total treatment effects, intermediate variables can of course provide useful information. For example, when we ask about the role of blood pressure in mediating the effect of treatment on recovery (as in Section 4) a whole set of mediation analytic techniques can be brought to bear on the question (e.g., VanderWeele 2009; Imai, Keele, and Yamamoto 2010; Pearl 2013) which aims to assess direct and indirect effects as formulated in Pearl (2001) and Robins and Greenland (1992). If, on the other hand, we ask questions about how the third variable (e.g., blood pressure) can help estimate treatment effects in the presence of unmeasured confounders, another set of tools is brought into consideration (see Pearl 1995). But when our query is "Which drug is more effective?" (assuming no unmeasured confounders, as in Figure 1(b)), the answer is unequivocal: "Ignore blood pressure."

Finally, I also agree with the spirit of Armistead's statement:

> Any conditioning on nontrivial third variable that produces Simpson Paradox should be carefully examined before either the aggregated or the disaggregated findings are accepted, regardless of whether the variable is thought to be causal.

I must point out, however, that we can do better than "carefully examine" the third variable. Modern tools of causal analysis now permit us to determine mathematically whether the aggregated or disaggregated findings should be accepted.[9] Specifically, in the blood-pressure example, mathematical analysis dictates that the aggregated findings give the correct answer to our specific research question, which is precisely what "careful examination" aims to accomplish. Armistead is correct in stating that this holds regardless of whether one categorizes "blood pressure" as causal or noncausal variable; what matters is the causal relationships of the third variable to other variables in the analysis, as portrayed in the diagram. Indeed, in Figure 1(c), for example, the third variable Z is not affected by the treatment, and still, it should not be controlled for; the aggregated finding should be accepted.

[9] Expressions such as "should be carefully examined" were used by statisticians in the precausal era to convey helplessness in handling causal questions.

4. CONCLUSIONS

I hope that playing the multistage Simpson's guessing game (Figure 3) would convince readers that we now understand most of the intricacies of Simpson's paradox, and we can safely title it "resolved."

REFERENCES

Agresti, A. (1983), "Fallacies, Statistical," in *Encyclopedia of Statistical Science* (vol. 3), eds. S. Kotz and N. Johnson, New York: Wiley, pp. 24–28. [8]

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. Available at *http://www.ete-online.com/content/5/1/5*. [8]

Berkson, J. (1946), "Limitations of the Application of Fourfold Table Analysis to Hospital Data," *Biometrics Bulletin*, 2, 47–53. [11]

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press. [8]

Blyth, C. (1972), "On Simpson's Paradox and the Sure-Thing Principle," *Journal of the American Statistical Association*, 67, 364–366. [8,9]

Cohen, M., and Nagel, E. (1934), *An Introduction to Logic and the Scientific Method*, New York: Harcourt, Brace and Company. [8]

Hernán, M., Clayton, D., and Keiding, N. (2011), "The Simpson's Paradox Unraveled," *International Journal of Epidemiology*, 40, 780–785. DOI: 10.1093/ije/dyr041. [xxxx]

Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71. [12]

Lauritzen, S. (1996), *Graphical Models* (reprinted 2004 with corrections), Oxford: Clarendon Press. [10]

Lindley, D., and Novick, M. (1981), "The Role of Exchangeability in Inference," *The Annals of Statistics*, 9, 45–58. [8]

Meek, C., and Glymour, C. (1994), "Conditioning and Intervening," *British Journal of Philosophy Science*, 45, 1001–1021. [xxxx]

Novick, M. (1983), "The Centrality of Lord's Paradox and Exchangeability for all Statistical Inference," in *Principles of Modern Psychological Measurement*, eds. H. Wainer and S. Messick, Hillsdale, NJ: Erlbaum, pp. 41–53. [9]

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann. [10]

——— (1993), "Comment: Graphical Models, Causality, and Intervention," *Statistical Science*, 8, 266–269. [9]

——— (1995), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–710. [12]

——— (2001), "Direct and Indirect Effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, pp. 411–420. [12]

——— (2009), *Causality: Models, Reasoning, and Inference* (2nd ed.), New York: Cambridge University Press. [8,10,11]

——— (2013), "Interpretation and Identification of Causal Mediation," *Psychological Methods*. [12]

Pearl, J., and Paz, A. (2013), "Confounding Equivalence in Causal Inference," Technical Report no. R-343w, Department of Computer Science, University of California, Los Angeles, CA. Revised and submitted, October 2013, available at *http://ftp.cs.ucla.edu/pub/stat_ser/r343w.pdf*. [11]

Pearson, K., Lee, A., and Bramley-Moore, L. (1899), "Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses," *Philosophical Transactions of the Royal Society of London*, Series A, 192, 257–330. [8]

Robins, J., and Greenland, S. (1992), "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155. [12]

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*, Corvallis, OR: AUAI, pp. 527–536. [11]

Simpson, E. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society,* Series B, 13, 238–241. [11]

VanderWeele, T. (2009), "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26. [12]

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer. [8]

Whittemore, A. (1978), "Collapsibility of Multidimensional Contingency Tables," *Journal of the Royal Statistical Society,* Series B, 40, 328–340. [8]

Yule, G. (1903), "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, 2, 121–134. [8]

Comment

Ronald CHRISTENSEN

I discuss predicting outcomes and the roles of causation and sampling design.

KEY WORDS: Causal models; Logistic; Logit; Loglinear; Prediction.

1. INTRODUCTION

In Dr. Armistead's examination of Simpson's paradox, there are three medical (agricultural) variables: an outcome variable recovery (yield), and two other variables: treatment (color) and one, but not both, of sex or blood pressure [BP] (height). Note that BP is assumed to be measured after treatments have been applied. The data are reproduced in Table 1. Simpson's paradox is that the treatment outperforms the control in the combined table which contradicts both the male and female tables.

Although I agree with the author that the data may have other uses, I will focus on predicting outcomes as well as the roles of causation and sampling design. For these data and the medical interpretations, one hopes to be in the population that recovers most frequently, and one makes choices that are consistent with that goal. With sex as the third medical variable, one hopes to be male, but that is not a choice, and regardless of sex, one chooses the control rather than the treatment. With BP as the third variable, one hopes to be in the normal group and chooses the control. However, in this medical version of Simpson's paradox, if one finds they are in the low BP group, a person would be well advised to switch to the treatment in the hope that it might put them into the normal group. In the agriculture version of the data, after choosing to plant black seeds (medical: control), and discovering that the plant is short (medical: low BP), one cannot go back and change the seed to being white in the hope that it becomes tall.

You can only make predictive choices based on the variables that are observed at the time the choice must be made. If predictive information is generally available but currently unobserved for the case to be predicted, it is wise to base decisions on an appropriate prior distribution for those unobserved variables. In other words, use an aggregated table that aggregates using the prior weights for the unobserved variables. Dr. Armistead illustrated this sort of aggregation for the observed variable sex using 50/50 weights. Weighting is discussed in much more detail in Section 3.

In the medical examples, it remains an article of faith that results on a new patient will be represented by the results of the data, that is, that the new patient is from the same population from which the data were sampled. Are patients assigned treatments? Assigning treatments creates two subpopulations to consider. Or do patients choose their treatments? Some treatments may be much more palatable to males than females. In these data, males got the treatment at a rate of three to one, whereas women got the control at a rate of three to one. Less faith seems needed in the agricultural example, only that the new plant is from the same populations sampled for the data.

2. CAUSATION

Christensen (1997, p. 212) argued somewhat controversially—see Spirtes, Glymour, and Scheines (2000)—that causation cannot be inferred from data analysis. Of course, given a collection of causal models, data analysis can help determine the better models.

In the medical version of the paradox relating recovery, an assigned treatment, and BP there are three self-evident causal models: treatment causes both recovery and BP.

Ronald Christensen, Department of Mathematics and Statistics, University of New Mexico, Mexico (E-mail: *fletcher@stat.unm.edu*). I would like to thank Joe Cavanaugh, who acted as editor on this discussion, for his valuable comments. Also, I would like to dedicate this discussion to Dennis Lindley who recently passed away.

 DOI: 10.1080/00031305.2014.876832