This article was downloaded by: [68.35.66.30]
On: 04 March 2014, At: 07:33
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

The American Statistician
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/utas20

Comment

Ronald Christensen [a]
[a] Department of Mathematics and Statistics , University of New Mexico , Mexico
Published online: 21 Feb 2014.

To cite this article: Ronald Christensen (2014) Comment, The American Statistician, 68:1, 13-17, DOI: 10.1080/00031305.2014.876832

To link to this article: http://dx.doi.org/10.1080/00031305.2014.876832

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://amstat.tandfonline.com/page/terms-and-conditions

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.

Downloaded by [68.35.66.30] at 07:33 04 March 2014

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.

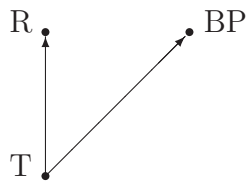
In the medical examples, it remains an article of faith that results on a new patient will be represented by the results of the data, that is, that the new patient is from the same population from which the data were sampled. Are patients assigned treatments? Assigning treatments creates two subpopulations 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

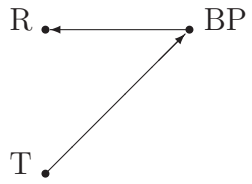
Christensen (1997, p. 212) argued somewhat controversially—see Spirtes, Glymour, and Scheines (2000)—that causation cannot be inferred from data analysis. Of course, given a collection of causal models, data analysis can help determine the better models.

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.

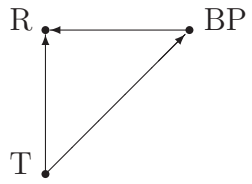
Ronald Christensen, Department of Mathematics and Statistics, University of New Mexico, Mexico (E-mail: *fletcher@stat.unm.edu*). I would like to thank Joe Cavanaugh, who acted as editor on this discussion, for his valuable comments. Also, I would like to dedicate this discussion to Dennis Lindley who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



Treatment causes BP which causes recovery.



And treatment is a direct cause of both recovery and BP, but BP is also a direct cause of recovery.



The third of these models is consistent with the treatment having either a direct negative or positive effect on recovery but, through the mechanism of the treatment normalizing BP, which helps recovery, perhaps positive overall and indirect treatment effects.

These causal models are consistent with log-linear models that, respectively, determine that recovery and BP are independent given treatment (logit/logistic model with treatment effect only), that treatment and recovery are independent given BP (logit/logistic model with BP effect only), and models that either include all two-factor interactions (logit/logistic model with treatment and BP main effects only) or include the three-factor interaction making it the saturated model (logit/logistic model with treatment and BP interaction). Nonetheless, finding the best-fitting model (among the few models being considered) is a far cry from determining that causation must follow because the model fits best. In particular, the fitted log-linear model would be as happy with recovery causing treatment as with treatment causing recovery. For example, the second causal graph with treatment causing BP which causes recovery suggests the causal decomposition

$$\Pr(R, T, \mathrm{BP}) = \Pr(R|\mathrm{BP}) \Pr(\mathrm{BP}|T) \Pr(T).$$

But the last two terms on the right merely define the joint distribution of BP and treatment, so this is probabilistically indistinguishable from

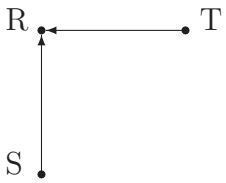
$$\Pr(R, T, \mathrm{BP}) = \Pr(R|\mathrm{BP}) \Pr(T|\mathrm{BP}) \Pr(\mathrm{BP}),$$

which is recovery and treatment independent given BP and what the log-linear model is traditionally described as fitting. Moreover, these are also indistinguishable from

$$\Pr(R, T, \mathrm{BP}) = \Pr(R) \Pr(\mathrm{BP}|R) \Pr(T|\mathrm{BP}),$$

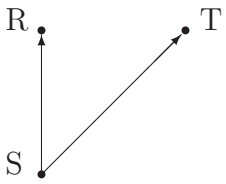
which are the probabilities suggested by reversing the causations.

The agricultural example works just like the BP example but, interestingly, in the medical examples the obvious causal models are different when we replace BP with sex. If treatments were assigned independently of sex, we might have the model



but in our data there is clearly a relationship between sex and treatment, regardless of whether treatments were assigned to people or if people got to choose their treatment. Moreover, this directed graph does not correspond to standard methods for modeling response variables in which one conditions on the predictor variables, here both treatment and sex; see Aldrich (2005). In particular, logit/logistic models are equivalent to log-linear models that contain the (highest order) interaction among all predictor variables, see Christensen (1997), but this graph excludes that sex–treatment interaction.

The two more viable causal models involving sex are



and

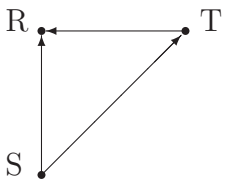


Table 1. Simpson's Paradox data.

	Recovery (Yield)		
Combined	+ (High)	− (Low)	Total
Treatment (White)	20	20	40
Control (Black)	16	24	40
Total	36	44	80
	Recovery (Yield)		
Male/Normal BP (Tall)	+ (High)	− (Low)	Total
Treatment (White)	18	12	30
Control (Black)	7	3	10
Total	25	15	40
	Recovery (Yield)		
Female/Low BP (Short)	+ (High)	− (Low)	Total
Treatment (White)	2	8	10
Control (Black)	9	21	30
Total	11	29	40

Downloaded by [68.35.66.30] at 07:33 04 March 2014

both of which are analogous to models that substitute BP for sex. Baker (2013) discussed causal models that involve unobserved variables.

Data analysis does not get at causation directly. The purpose of randomly assigning treatments to experimental material is that it provides a philosophical basis for inferring that the treatments cause whatever effect we may see. Randomization does not work perfectly, we can get bad randomizations, but the chance of repeated bad randomizations is small. Randomizing treatment assignment means that the levels of any confounding variables should be distributed about evenly between the treatments. In the medical data under a completely randomized design (CRD), we would have to be extremely unlucky to get a 3 to 1 split of males getting the treatment and another 3 to 1 split of the females getting the control. It could happen, but it would not happen very often. And if you thought that recovery was likely to vary a lot by sex, you would use a randomized block design blocking on sexes. Again, it would be an unusual randomization that gave 3 to 1 treatment splits within blocks, but in the analysis of a block design that would not matter because the analysis *must* include block/sex effects, so Simpson's paradox should not raise its ugly head.

In fact, there is strong internal evidence that these data were collected by intentionally assigning treatments at a 3 to 1 rate within blocks (block–treatment counts are multiples of 10) so that the data collection scheme mandates an analysis including block effects. These data have a clearly detectable block by treatment interaction, something that is only detectable because each block contains more subjects than there are treatments. In such cases, we must decide whether the block by treatment interaction is of interest in its own right or whether block by treatment interaction determines the appropriate error term to evaluate whether treatment main effects are significant. (If evidence for main effects is not so blatant that it overwhelms any block–treatment interaction we should not declare main effects.) Neither the BP nor height scenarios from the article allow blocking because the variables are only determined after the treatments are assigned (although one could determine BP prior to treatment assignment). And, although one can assign treatments without randomization, any inference of causation from these data requires random assignment of treatments to experimental material, even if at a 3 to 1 ratio.

The key point is that randomization guards against the existence of other confounding variables that could skew the conclusions like sex skews them. As the author demonstrates, there would have been no problem with the marginal table if the sexes were evenly distributed between treatments and in a CRD the randomization would likely have made that approximately true. We should note as did Cox (1958, Sec. 4.2), more than 50 years ago, that the causal validity of a randomized experiment can be compromised by the inclusion in the data analysis of inappropriate covariates that are themselves affected by the treatments. Thus, neither (post-treatment) BP nor height would be allowed in analyzing causal effects from these data if they came from a randomized experiment on treatments (seed colors). Nonetheless, as the author points out, relationships to such variables can still be of interest.

3. PREDICTION AND SAMPLING

At the cost of getting a little more technical about predictive distributions and sampling distributions, I think the issues clarify further.

Two things seem self-evident: that we should collect as much data as we can afford and that we should use as much data as possible to make predictions. The other information that is particularly germane is identifying how the data were collected. We identify y with recovery (yield), x_1 with treatment (color), and x_2 with sex or BP (height). Let $f(y, x_1, x_2)$ denote the joint density or probability mass function. We would like to use $f(y|x_1, x_2)$ as the basis for predictions of y because it uses the most predictive information by conditioning on known values for both x_1 and x_2, that is, we would like to use probability estimates from the disaggregated table. Unfortunately, only the scenario with x_2 being sex allows us to observe and condition on x_2 when choosing a treatment or seed color. In the other two scenarios, x_2 is not observed at the time a prediction must be made, so predictions must be based on $f(y|x_1)$ because the only predictive information we have is x_1, that is, we cannot use the probability estimates from the disaggregated table.

Now let us look at the data that can be collected. Internal evidence from Table 1 suggests that the data were collected as independent samples from the four treatment–sex populations, that is, they are samples from $f(y|x_1, x_2)$. The internal evidence referred to is that the sample size for every combination of (x_1, x_2) is a multiple of 10. From such a sample we can estimate the conditional probabilities $f(y|x_1, x_2)$ directly from the disaggregated table and make the obvious inferences. While this sampling scheme is most natural for x_2 as sex, one could also take samples from all four treatment–BP categories or all four color–height categories by picking the samples after the populations have manifested themselves.

A number of other possible sampling schemes could give rise to the disaggregated table. If we are assigning treatments to people, we must sample from a distribution that conditions on treatments x_1. Rather than samples that condition on both x_1 and x_2 we could sample from each treatment (color) group, that is, sample from $f(y, x_2|x_1,)$. If we are not assigning treatments we could take a random sample of the entire population, that is, sample from $f(y, x_1, x_2)$. For example, we could imagine a population of seeds, some of which are white and some of which are black. Or we could sample a collection of patients and note their recovery, treatment, and sex/BP. Alternatively, we could take independent samples from the two sexes, that is, sample $f(y, x_1|x_2)$. This is also viable in the other two scenarios. We could take a sample of patients from each BP group and see what treatment they received and their recovery status. Similarly, we could sample tall plants and see what color their seeds were and their yields. These four sampling schemes are all consistent with performing regression. Finally, we could sample from $f(x_1, x_2|y)$ which is the traditional discrimination problem. For example, we can sample from recovered and unrecovered patients to examine their treatment and sex/BP. Of course, there are other possible sampling schemes that condition on y but we consider only this one.

Downloaded by [68.35.66.30] at 07:33 04 March 2014

From any of the four regression sampling schemes, it is a simple matter to estimate $f(y|x_1, x_2)$. This amounts to using the conditional probability estimates from the disaggregated table. Estimating $f(y|x_1)$ from the four regression sampling schemes requires more thought, which is why using the aggregated table may cause problems.

If we sampled either the entire population $f(y, x_1, x_2)$ or sampled the different treatment (color) populations $f(y, x_2|x_1)$, the aggregated table gives direct information on

$$f(y_1, x_1) = \sum_{x_2} f(y, x_1, x_2)$$

and

$$f(y_1|x_1) = \sum_{x_2} f(y, x_2|x_1),$$

respectively. The former allows easy computation of probabilities conditional on x_1 and the latter is the end product. As mentioned, because both are just marginal distributions relative to the sampling distribution, the simple aggregated data provide appropriate estimates of these probabilities. In both the BP medical and agricultural scenarios, these are reasonable regression sampling schemes that lead to using the aggregated table when making predictions.

If we condition on x_2 in the sampling scheme, finding the predictive distribution $f(y|x_1)$ requires us to weight the samples using a distribution on x_2.

When sampling from treatment, sex groups as these data probably would have been, it takes some thought to imagine how you could sample from populations with both x_1 and x_2 fixed but not know x_2 when you want to make predictions. Perhaps measuring x_2 is very expensive. In any case, when sampling from $f(y|x_1, x_2)$,

$$f(y_1|x_1) = \sum_{x_2} f(y|x_1, x_2) f(x_2|x_1),$$

so before an aggregated table for x_1 and y would be useful for prediction, the samples need to be weighted by the conditional distribution of x_2 given x_1. This conditional distribution is not estimable when the data are sampled from $f(y|x_1, x_2)$, must be found outside the study, and can get quite complicated. In cases where x_1 is a randomly assigned treatment, it makes sense that the weighting function $f(x_2|x_1)$ would just be the marginal distribution $f(x_2)$. So, as illustrated by the author, assuming that males and females are equally likely in the population, the estimated recovery probabilities given each treatment and sex have to be given equal weights, rather than the 3 to 1 weights that are implicit in the aggregated table based on the disproportionate sampling rates that were used in data collection. Note that if $f(x_2|x_1)$ is defined by Pr(Male|Treatment) = 0.75 and Pr(Female|Control) = 0.75 we would get predictions consistent with the aggregated table. However, these numbers only agree with males choosing treatment and females choosing control at rates of 3 to 1, that is, Pr(Treatment|Male) = 0.75 and Pr(Control|Female) = 0.75, because the sex ratio is taken as 50/50.

Similarly, if sampling from each level of x_2 (sex, BP, height), we must weight by the marginal distribution of x_2,

$$f(y_1, x_1) = \sum_{x_2} f(y, x_1|x_2) f(x_2).$$

Again, the marginal distribution of x_2 must be known outside the current study. This sampling scheme allows, for example, treatments to depend on sex. For some reason, in these data 75% of men would have chosen the treatment, while 75% of women would have chosen the control. But to get a reasonable prediction based on choosing a treatment for a person of unknown sex, we again need to weight the treatment results by the proportions of women in the population rather than the proportion in the study. For these data, the population and study sex proportions are both about 50/50 so this aggregated table leads to appropriate conclusions for sex. We have no reason to believe that the population proportions of BP or height would be 50/50, so no reason to think that the aggregated table would be appropriate in those circumstances.

Finally, these sorts of computations are much more familiar when dealing with discrimination data. It is quite standard to argue that for discrimination data sampled from $f(x_1, x_2|y)$, we need to use Bayes theorem to compute

$$f(y|x_1, x_2) = \frac{f(x_1, x_2|y) f(y)}{\sum_y f(x_1, x_2|y) f(y)},$$

where we need to know the prevalences of the conditions, $f(y)$, from outside the study. However, in the current situation we really need the less standard computation

$$f(y|x_1) = \sum_{x_2} f(y|x_1, x_2) f(x_2|x_1)$$

or perhaps more simply compute

$$f(y, x_1) = \sum_{x_2} f(x_1, x_2|y) f(y)$$

before finding the conditional probabilities.

Perhaps I should also note that these issues do not usually arise when doing variable selection in a regression analysis. In such analyses, we often determine that, say, x_2 has no effect on $f(y|x_1, x_2)$ so we can perform regression using x_1 alone, that is, using $f(y|x_1) = f(y|x_1, x_2)$. Here, we are trying to predict using only x_1 when we know that x_2 is an important predictor variable.

4. FINAL THOUGHTS

I would like to thank Dr. Armistead for his stimulating contribution.

To me, a crucial distinction is to be made between predictive models and causative models. The examination of Simpson's paradox reinforces my belief that it is not safe to infer causation unless one has conducted a randomized experiment. Admittedly, the randomizations can go bad, which is one reason that replicating experiments is so important.

If the discussion of Simpson's paradox is not about causation and only about prediction, then the analysis takes a different bent. The default is to use all of the information available to

Downloaded by [68.35.66.30] at 07:33 04 March 2014

make predictions. However, if you do not know the height of a plant or the sex of a person or their blood pressure when deciding on a treatment, then you cannot use them. In such cases, the appropriate predictive procedure depends crucially on how the data were obtained. If predictor information becomes available later, and you can change treatments, you might want to do that.

While prediction is the ultimate goal of science, causation is the warm fuzzy. Causation can greatly simplify prediction and we like to think that good causative models provide the best predictions. But in the end, getting predictions correct is more important than imagining that we understand why things happen the way they do. While I admit that I am not an expert on the causal model literature, I am unfamiliar with any satisfactory way to infer causation other than performing randomized experiments. Sure, data analysis can help you choose between two or more causative models, but that is a far cry from inferring causation from data analysis. In fact, without knowing the sampling design, we cannot even be sure of making appropriate predictions from data analysis alone.

REFERENCES

Aldrich, J. (2005), "Fisher and Regression," *Statistical Science*, 20, 401–417. [14]

Baker, S. G. (2013), "Causal Inference, Probability Theory, and Graphical Insights," *Statistics in Medicine*, 32, 4319–4330. [15]

Christensen, R. (1997), *Log-Linear Models and Logistic Regression* (2nd ed.), New York: Springer. [13,14]

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley. [15]

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search* (2nd ed.), Cambridge: MIT Press. [13]

Downloaded by [68.35.66.30] at 07:33 04 March 2014

Comment: A Fruitful Resolution to Simpson's Paradox via Multiresolution Inference

Keli LIU and Xiao-Li MENG

Simpson's Paradox is really a Simple Paradox if one at all. Peeling away the paradox is as easy (or hard) as avoiding a comparison of apples and oranges, a concept requiring no mention of causality. We show how the commonly adopted notation has committed the gross-ery mistake of tagging unlike fruit with alike labels. Hence, the "fruitful" question to ask is not "Do we condition on the third variable?" but rather "Are two fruits, which appear similar, actually similar at their core?." We introduce the concept of *intrinsic* similarity to escape this bind. The notion of "core" depends on how deep one looks—the multi resolution inference framework provides a natural way to define intrinsic similarity at the resolution appropriate for the treatment. To harvest the fruits of this insight, we will need to estimate intrinsic similarity, which often results in an indirect conditioning on the "third variable." A ripening estimation theory shows that the standard treatment comparisons, unconditional or conditional on the third variable, are low hanging fruit but often rotten. We pose assumptions to pluck away higher-resolution (more conditional) comparisons—the multiresolution framework allows us to rigorously assess the price of these assumptions against the resulting yield. One such assessment gives us *Simpson's Warning: less conditioning is most likely to lead to serious bias when Simpson's Paradox appears*.

KEY WORDS: Bias-variance tradeoff; Principal stratification

1. THE SOURCE OF CONFUSIONS AND DEBATES

1.1 Comparing Apples and Oranges

Imagine Ms. Broken going to Dr. Heal to be treated for heart disease. A new treatment was made available to Dr. Heal, who also learned from a clinical trial that it can substantially outperform a standard treatment used as its control. However, its effectiveness depends on a patient's cholesterol level, which can also be altered significantly by the treatment. Therefore, to determine the appropriate treatment for Ms. Broken, Dr. Heal needs to know how trial subjects with cholesterol level similar to Ms. Broken's (say about 240 mg/dL) responded to the

Keli Liu (E-mail: *keli.liu25@gmail.com*) is A.B. Graduate, and Xiao-Li Meng (E-mail: *meng@stat.harvard.edu*) is Whipple V. N. Jones Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. The authors thank Dr. Armistead for an invigorating article that stimulated authors to take a higher resolution look at Simpson's Paradox. Jessica Hwang provided invaluable advice on organization and structure, as well as pointing out incoherencies and inconsistencies. Of course, the remaining incoherencies and inconsistencies are entirely of the authors. The authors also thank the NSF for partial financial support, and *TAS* editor, Ronald Christensen, and journal manager, Eric Sampson, for their saintly patience amid growing despair. During the preparation of this discussion, we learned the sad news of Dennis Lindley's passing. Without his work, our understanding of this and many other topics would not be as rich today. We therefore dedicate this article in his memory.

© 2014 American Statistical Association DOI: 10.1080/00031305.2014.876842