TECHNICAL REPORT
R-436
July 2016

Lord's Paradox Revisited – (Oh Lord! Kumbaya!)

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

July 14, 2016

Abstract

Among the many peculiarities that were dubbed "paradoxes" by well meaning statisticians, the one reported by Frederic M. Lord in 1967 has earned a special status. Although it can be viewed, formally, as a version of Simpson's paradox (Arah, 2008; Tu et al., 2008; Pearl, 2014b) its reputation has gone much worse. Unlike Simpson's reversal, Lord's is easier to state, harder to disentangle (Wainer and Brown, 2007) and, for some reason, it has been lingering for almost four decades, under several interpretations and re-interpretations (Holland and Rubin, 1983), and it keeps coming up in new situations and under new lights (van Breukelen, 2013; Senn, 2006; Eriksson and Häggström, 2014). Most peculiar yet, while some of its variants has received a satisfactory resolution (Glymour, 2006; Hernández-Díaz et al., 2006), the original version presented by Lord, to the best of my knowledge, has not been given a proper treatment, not to mention a resolution.

The purpose of this paper is to trace back Lord's paradox from its original formulation, resolve it using modern tools of causal analysis, explain why it resisted prior attempts at resolution and, finally, address the general methodological issue of whether adjustments for pre-existing conditions is justified in group comparison applications.

1 Lord's original dilemma

Any attempt to describe Lord's paradox in words other than those used by Lord himself can only do injustice to the clarity and freshness with which it was first enunciated in 1967. We will begin therefore by listening to Lord's own words.

> "A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded.

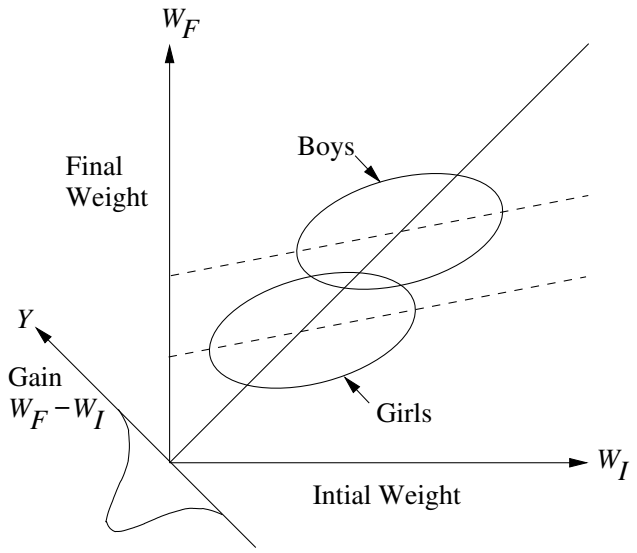

Figure 1: Lord's method of displaying no change in average gain ($W_F - W_I$) co-habitating with an increase in adjusted weight.

At the end of the school year, the data are independently examined by two statisticians. Both statisticians divide the students according to sex. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds these to be identical. On further investigation, he finds that the frequency distribution of weight for the girls at the end of the year is actually the same as it was at the beginning.

He finds the same to be true for the boys. Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year.

The whole situation is shown by the solid lines in the diagram [Fig. 1]. Here the two ellipses represent separate scatter-plots for the boys and the girls. The frequency distributions of initial weight are indicated at the top of the diagram and the identical distributions of final weight are indicated on the left side. People falling on the solid 45° line through the origin are people whose initial and final weight are identical. The fact that the center of each ellipse lies on this 45° line represents the fact that there is no mean gain for either sex.

The first statistician concludes that as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.

The second statistician, working independently, decides to do an analysis of covariance. After some necessary preliminaries, he determines that the slope of the regression line of final weight on initial weight is essentially the same for

the two sexes. This is fortunate since it makes possible a fruitful comparison of the intercepts of the regression lines. (The two regression lines are shown in the diagram as dotted lines. The figure is accurately drawn, so that these regression lines have the appropriate mathematical relationships to the ellipses and to the 45° line through the origin.) He finds that the difference between the intercepts is statistically highly significant.

The second statistician concludes, as is customary in such cases, that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes. When pressed to explain the meaning of this conclusion in more precise terms, he points out the following: If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distributions of initial weight, the relative position of the regression lines shows that the subgroup of boys is going to gain substantially more during the year than the subgroup of girls.

The college dietician is having some difficulty reconciling the conclusions of the two statisticians. The first statistician asserts that there is no evidence of any trend or change during the year for either boys or girls, and consequently, a fortiori, no evidence of a differential change between the sexes. The data clearly support the first statistician since the distribution of weight has not changed for either sex.

The second statistician insists that wherever boys and girls start with the same initial weight, it is visually (as well as statistically) obvious from the scatter-plot that the subgroup of boys gains more than the subgroup of girls.

It seems to the present writer that if the dietician had only one statistician, she would reach very different conclusions depending on whether this were the first statistician or the second. On the other hand, granted the usual linearity assumptions of the analysis of covariance, the conclusions of each statistician are visibly correct.

This paradox seems to impose a difficult interpretative task on those who wish to make similar studies of preformed groups. It seems likely that confused interpretations may arise from such studies.

What is the "explanation" of the paradox? There are as many different explanations as there are explainers.

In the writer's opinion, the explanation is that with the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups. The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data." (Lord, 1967)

These pessimistic words conclude Lord's narrative, and became a challenge to almost half a century of speculations and interpretations. Most worthy of attention is his counterfactual

definition of the research problem: "The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences."

2 Interpretation

Before attempting to cast Lord's story in a formal setting, let us examine whether his dilemma is expressed convincingly.

Since no description is given nor data taken under the old diet, the dilemma faced cannot focus on a comparison between the two diets, old and new. Rather, the new diet must be taken as a given condition, which, together with time, metabolism and natural growth has brought about weight changes in some individuals, from their initial weight (W_I) in September, to their final weight (W_F) in June. The research question at hand is whether the weight change process (under a fixed diet condition) is the same for the two sexes. In other words, the question is whether the distinct metabolism of boys has a different effect on their growth pattern than that of girls, under the given diet. Indeed, differential gain is the main concern of both statisticians: the first concludes that "there is no evidence of any differential effect on the two sexes," and the second insists that "whether boys and girls start with the same initial weight, ... the subgroup of boys gains more that the subgroup of girls." The issue of assessing differential gain "under the same initial conditions" is further emphasized in Lord's last paragraph, stating: "The researcher wants to know how the groups would have compared if there were no preexisting uncontrolled differences." Here the use of the counterfactual expression "if there were no preexisting differences" leaves no doubt that it is the effect of gender on weight gain that is the center of investigation while diet, since it is common to all subjects, should be treated as a fixed background condition.

With this understanding of the research question, what is the difference between the two statisticians? Both were asked to determine if there is a differential gain among the sexes but they came back with a different answer. Statistician-1 simply compared the weight gain distributions of the two groups and concluded that there is no change. The perfect overlap of the two ellipses on the 45° line indicates that there is no difference in growth rate of the two sexes.

Statistician-2 however noticed that the initial weight of boys is higher (on average) than that of girls and, moreover, since the difference in initial weight can plausibly be attributed to their gender difference, he decides to "make proper allowance" for this difference and adjust for W_I, so as to compare the groups on the basis of gender alone. Here, he finds that Boys gain more than girl in every stratum of W_I so, naturally, he concludes that boys gain more than girls on the average, contrary to statistician-1.

Thus, the paradox which we need to address is: Why should a greater weight gain (for men) which is found in every stratum of the initial weight W_I suddenly disappear when averaged over the group as a whole. In other words, we expect the finding of statistician-2 to constrain the finding of statistician-1. We feel that they should comply with the "Sure Thing Principle" (Savage, 1962; Pearl, 2016), which states (loosely): "A relation that holds in every subpopulation should not disappear or reverse sign when applied to the population as a whole." Violation of this principle is behind Simpson's paradox ("good for men, good for women yet bad for people") and it is this violation that must have triggered Lord's

astonishment as to why the two statisticians do not arrive at the same conclusion.

Note that this astonishment haunts us regardless of what takes place under the old diet; the data generated under the new diet (Figure 1) is sufficient to make us wonder why it is that generalizing what statistician-2 finds in every stratum of W_I (i.e., and increase gain for males) contradicts what statistician-1 finds in the population as a whole (i.e., no increase overall).

The resolution of the paradox is the same as the resolution of Simpson's paradox: The sure thing principle does not forbid reversal (or disappearance) of local associations upon aggregation, it forbids only reversal of causal effects when the subpopulations remains of the same size. In our case, the subpopulations characterized by each stratum of W_I do nor remain constant as we move from males to females, girls populate the underweight strata much more than boys.

The clearest way to see that association reversal should not betray our intuition (nor the sure thing principle) is to view gender as the treatment variable and examine its effect on weight gain.

With this understanding of the research question, we are facing a mediation problem in which the initial weight mediates the causal process between gender and the final weight. The first statistician estimated the *total effect* (of gender on gain) while the second statisticians estimated the *direct effect*, adjusting for the mediator, W_I.[1]

Put in these terms, it should come as no surprise that the two statisticians came up with different, but hardly contradictory, answers. Cases where total and direct effects differ in sign and magnitude are commonplace. For example, we are not at all surprised when smallpox inoculation carries risks of fatal reaction, yet reduces overall mortality by iradicating smallpox. The direct effect (fatal reaction) in this case is negative for every stratum of the population, yet the total effect (on mortality) is positive for the population as a whole.

Thus, Lord's pessimistic conclusions were rather premature. It is not the case that "there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups." On the contrary, such procedures, though not available in Lord's time, are now well developed in the causal mediation literature (Robins and Greenland, 1992; Pearl, 2001; Imai et al., 2010; Valeri and VanderWeele, 2013). They require only that researchers specify in advance whether it is the direct or total effect that is the target of their investigations. Both statisticians were in fact correct, though each estimated a different effect. statistician-1 aimed at estimating the total effect (of gender on weight gain) and, based on the data available properly concluded that there is no gender difference. The second statistician aimed at estimating the direct effect of gender on weight gain, unmediated by the initial weight and, after properly adjusting for the initial weight (i.e., the mediator) rightly concluded that there is significant gender difference, as seen through the displaced ellipses.

In the next section we provide a formal analysis for these two research questions.

[1]Readers who feel uncomfortable treating gender as a cause can think of the make up of gender-specific hormones as the causal variable; it causes differences in initial weight, and may also have direct effect on how a student responds to the new diet.

3 The paradox in a formal setting

The diagram in Fig. 2 describes Lord's dilemma as interpreted in the previous section. In

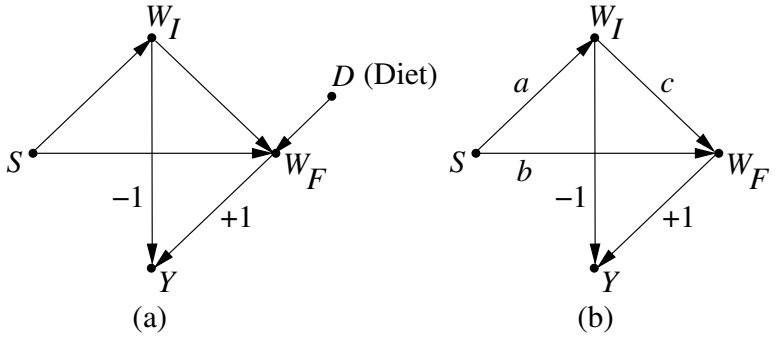

Figure 2: (a) Lord's model, showing Initial Weight (W_I) as a mediator between Sex (S) and Final Weight (W_F). (b) A linear version of (a).

this model S stands for Sex, W_I for the initial weight, W_F for the final weight and Y for the gain $W_F - W_I$. As the diagram shows, the initial weight W_I is affected by Sex and affects the final weight. It is thus a mediator between S and W_F as well as between S and the gain Y.

Assuming no confounding,[2] the nonparametric mediation model for Fig. 2(a) lends itself to simple analysis; both the total effect and direct effect are estimable from the data (Pearl, 2001). In particular, the total effect is given by the regression

$$TE = E(Y|S = 1) - E(Y|S = 0),$$

while the direct effect is given by

$$DE = \sum_w [E(Y|S = 1, W_I = w) - E(Y|S = 0, W_I = w)]P(W_I = w|S = 0).$$

Here we take $S = 1$ to represent boys and $S = 0$ to represent girls.[3]

Clearly these two expressions are quite different; there is no wonder therefore that they give different estimates. In Lord's example, the total effect is zero, as confirmed by statistician-1's observation that the two ellipses map into identical projections onto the 45° line, and the direct effect (with the baseline W_I as mediator) is non-zero, as seen by statistician-2, who observed the displaced ellipses for every stratum $W_I = w$.

An algebraic way of seeing how these results can come about is provided by the linear version of the model, shown in Fig. 2(b). Assuming standardized variables, the total effect is

[2]Since Sex is an exogenous variable, it acts "as if randomized," and its total effect is not confounded; it can be estimated by regression. However, the $W_I \to W_F$ relationship may be confounded by unobserved common causes of the two, which might distort the direct effect. We discuss this situation in Section 6; here we assume no such confounding.

[3]Readers will recognize the expression for DE as the "Natural Direct Effect" (Pearl, 2001) or the "Mediation Formula" which has become standard in mediation analysis (VanderWeele, 2009; Imai et al., 2010). (See Pearl (2014a) for identification conditions.)

given by the sum of the products of all coefficients along paths from S to Y (Wright, 1921; Pearl, 2013),

$$TE = (b + ac) - a = b - a(1 - c)$$

while the direct effect skips the paths going through W_I, and gives

$$DE = b.$$

The observed condition of zero total effect can easily be realized by setting $b = a(1-c)$, which accounts for the observations shown in Fig. 1. We see that the total effect TE vanishes due to cancelation of the three paths leading from S to Y; the direct effect is positive (b), while the indirect effect is equal and negative, resulting in zero total effect. Translated, whereas on average a boy gains more than a girl of equal initial weight, the fact that sex differences produce more heavy-weight boys than girls and that we subtract a portion of this difference, renders the overall gain for boys equal to that of girls.

4 Other versions of Lord's paradox

Early efforts to resolve Lord's paradox were made by Bock (1975, pp. 490–496), Judd and Kenny (1981b); Cox and McCullagh (1982), and Holland and Rubin (1983). Since no data was given on the old-diet, authors had to assume a model of weight gain under old-diet conditions and concluded, almost uniformly, that both statisticians were in fact correct, depending on the model assumed and on the precise questions that the statisticians attempted to answer. Bock, for example, sees no contradiction between the two statisticians. The first statistician asks: "Is there a difference in the average gain in weight of the population?" and correctly answered: "No!" The second statistician asks: "Is a man expected to show a greater weight gain than a woman, given that they are initially of the same weight?" and answers it correctly: "Yes!" (Bock, 1975, p. 491). Bock does not explain why the two conclusions are noncontradictory given the the first is merely generalization of the second.

Cox and McCullagh (1982) computed the causal effect of the new diet by assuming that, under the old diet, the final weight of every individual will remain the same as the initial weight. Accordingly, they found that statistician-1 is correct, the average causal effect (ACE) of the new diet on weight gain is zero for both men and women. Based on the same model, they found that statistician-2 is also correct, though he simply asks a different question, concerning the behavior of individual units within each population. Here statistician-2 finds that individual units are affected differently; initially overweight individuals tend to lose weight, and initially underweight individuals tend to gain weight. Naturally, then, comparing boys and girls at the same initial weight would show boys losing more weight than girls. Again, what Cox and McCullagh left unanswered is why the two findings – differential gain on every stratum and equal gain on the average – should not contradict the "sure thing" principle.

Holland and Rubin assumed several different models for the old-diet and showed that, in contrast to the Cox and McCullagh's model, the gender specific causal effects of the diet may be non-zero for both men and women, and their difference can be either positive or negative depending on the parameters of the assumed model. Thus, conclude Holland

and Rubin, neither statistician is correct or incorrect; it all depends on which model one assumes for the old diet weight gain. What Holland and Rubin did not explain is what in the new-diet data gave Lord's the unmistaken impression that statisticians 1 and 2 reach conflicting conclusions, namely, why their findings should not be constrained by the Sure Thing Principle.

Another question left unanswered by early interpreters is Lord's appeal for a general strategy of "allowing" for initial group differences. "The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences." In other words, is there a general criterion for deciding whether controlling for pre-treatment differences is a valid thing to do, in case we wish to compare group behavior that is free from the influence of those differences.

Such a general criterion is provided by the graphical analysis presented in the previous section. The criterion coincides with the answer to the question of whether adjustment for covariates (in our case, W_I) is appropriate for estimating total and direct effects. It is based on the graph structure alone, free of parametric assumptions that renders the analysis of Holland and Rubin undecisive.

Holland and Rubin did not attempt to interpret the problem in terms of the effect of gender, as we did in the previous section, because gender, being unmanipulable, cannot have a causal effect according to Holland and Rubin's doctrine of "no causation without manipulation" (Holland, 1986). let us apply it to a model proposed by Wainer and Brown (2007), where the target quantity is the effect of diet, not of gender. Wainer and Brown simplified Lord's original problem and interpreted the two ellipses of Fig. 1 to represent two different diets, or two dining halls, each serving a different diet. They further removed gender from consideration and obtained the two data sets seen in Fig. 3 [their Figure 9]. Since the choice of dining tables is manipulable, causal effects are well defined, and they presented Lord's dilemma as choosing between two methods of estimating the causal effect of dining room on weight gain. In their words:

> "The first statistician calculated the difference between each student's weight in June and in September, and found that the average weight gain in each dining room was zero. This result is depicted graphically in Fig. 3 [their Figure 9]. with the bivariate dispersion within each dining hall shown as an oval. Note how the distribution of differences is symmetric around the 45° line (the principal axis for both groups) that is shown graphically by the distribution curve reflecting the statistician's findings of no differential effect of dining room.
>
> The second statistician covaried out each student's weight in September from his/her weight in June and discovered that the average weight gain was greater in Dining Room B than in Dining Room A. This result is depicted graphically in Fig. 4 [their Figure 10]. In this figure the two drawn-in lines represent the regression lines associated with each dining hall. They are not the same as the principal axes because the relationship between September and June is not perfect. Note how the distribution of adjusted weights in June is symmetric around each of the two different regression lines.[4] From this result the second

[4]The regression line is the line along which the distribution of final weight, achieves its maximum value,

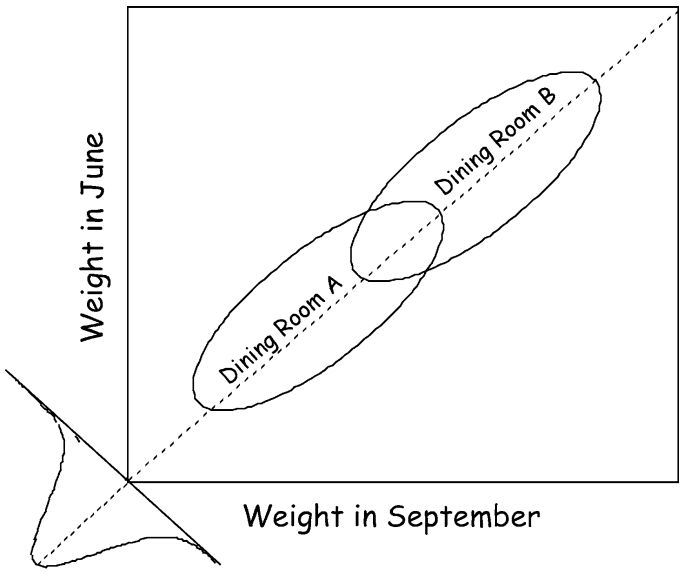

Figure 3: A scatter plot of a simplified Lord's Paradox showing the bivariate distribution of weights in two dining rooms at the beginning and end of each year [after Wainer and Brown, 2007].

> statistician concluded that there was a differential effect of dining room, and that the average size of the effect was the distance between the two regression lines.
>
> So, the first statistician concluded that there was no effect of dining room on weight gain and the second concluded there was. Who was right? Should we use change scores or an analysis of covariance? To decide which of Lord's two statistician's had the correct answer requires that we make clear exactly what was the question being asked. The most plausible question is causal, 'What was the causal effect of eating in Dining Room B?' " (Wainer and Brown, 2007)

Wainer and Brown's model is depicted in Fig. 5. Here, the initial weight is no longer treatment dependent for it was measured prior to treatment. It is in fact a confounder since, as shown in the data of Fig. 3 [their Figure 9], overweight students seem more inclined to choose Dining Room B, compared with underweight students. So, W_I affects both diet (D) and final weight (W).

It is clear from the graph of Fig. 5 that, regardless of whether one aims at estimating the effect of diet on the final weight W_F or on the weight gain (Y) adjustment for the initial weight W_I is necessary. Thus, statistician-2, who adjusted for W_I (ANCOVA) was correct, while statistician-1, who was charmed by the equality of average weight gain under the two diets was flatly wrong. This equality reflects no change in expected weight gain predicated upon *finding* a subject in Dining Room A as compared to B; it does not represent equality of gains *due* to a change from Dining Room A to dining room B. Confounders need to

for any given initial weight.

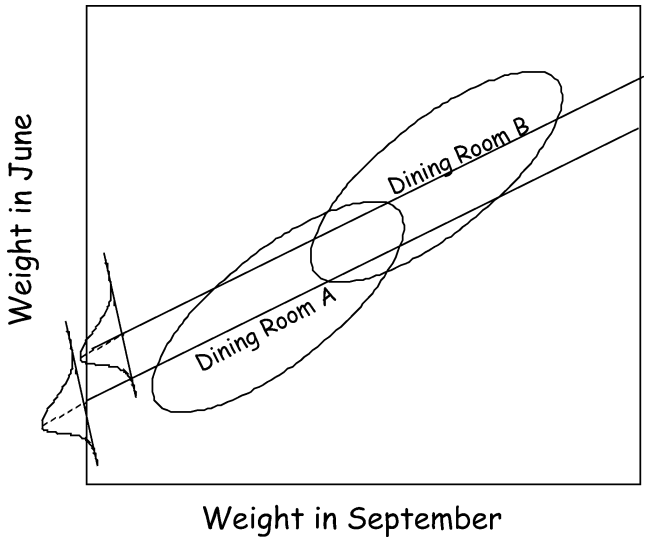

Figure 4: A graphical depiction of Lord's Paradox showing the bivariate distribution of weights in two dining rooms at the beginning and end of each year augmented by the regression lines for each group [after Wainer and Brown, 2007].

be "controlled for" when causal effects are estimated, and failure to do so leads to biased results. The right answer, therefore, lies with statistician-2, who concluded that diet A led to significantly more gain in weight than diet B when proper allowance is made for differences in initial weight between the two groups. This also explains why the Sure Thing Principle need not constrain the predictions of the two statistician; the principle applies to causal effects, not to statistical predictions (Pearl, 2016).

Interestingly, Wainer and Brown did not reach this conclusion. Instead, they concluded that the two statisticians were right, but made different assumptions. In their words:

> "To draw his conclusion the first statistician makes the implicit assumption that a student's control diet (whatever that might be) would have left the student with the same weight in June as he had in September. This is entirely untestable. The second statistician's conclusions are dependent on an allied, but different, untestable assumption. This assumption is that the student's weight in June, under the unadministered control condition, is a linear function of his weight in September. Further, that the same linear function must apply to all students in the same dining room."

I differ from Wainer and Brown in this conclusion. There is no need for the assumption of linearity to justify the correctness of statistician-2's insistence on using ANCOVA. Simultaneously, no assumption whatsoever would justify statistician-1 conclusion. Failure to control for confounding cannot be remedied by linearity, and proper control for confounder works both in linear and nonlinear models.

It is worth re-emphasizing at this point that our analysis relies, of course, upon the assumption of no unobserved confounders. When latent confounders are present, the ma-

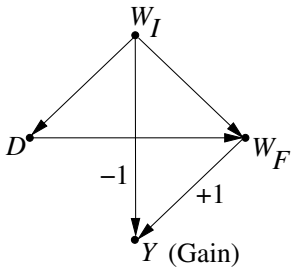

Figure 5: Graphical representation of Wainer and Brown's scenario in which the initial weight (W_I) is a determiner of diet (D), and the effect of Diet on gain requires an adjustment for W_I.

chinery of *do*-calculus (Pearl, 1994; Shpitser and Pearl, 2008) need be invoked to decide if the target effects are estimable or not. If not, then both statisticians are wrong, none of the two methods would result in unbiased estimate, and Lord's despair is perhaps justified: "The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the bases of available data."

However, the need to invoke causal assumptions, beyond the available data (e.g., no unmeasured confounding) applies to ALL tasks of causal inference (in observational studies), so there is nothing special to Lord's paradox. The unique challenge that Lord's paradox presented to the research community was to decide, from a rudimentary qualitative model of reality, whether allowance for preexisting differences should be made and, if so, how. We have seen that in the case of Lord's original story (Fig. 1) as well as in the dining rooms variant of the story (Fig. 3) such determination could be made using plausible qualitative models, without making any assumptions about the functional form of the relationship between a treatment and its outcomes.[5]

In the first story, both statisticians were right, each aiming at a different effect. In the second story, one was right (ANCOVA) and one was wrong. But in no case did we face a predicament like the one that triggered Lord's curiosity: two seemingly legitimate methods giving two different answers to the same research question. Lord gave in to the clash, and declared surrender. But he shouldn't have; whether we can estimate a given effect or not (for a given scenario) is a mathematical question with a yes/no answer, and should not be shaken by a clash of intuitions.

[5]In all fairness to Holland and Rubin, one should mention that the facility to make this determination (i.e., for any qualitative model, regardless how complex), was not available in 1983; it was developed a decade later and was kept relatively unknown in potential outcome circles (Pearl, 1993; Rubin, 2004; Pearl, 2009b; Rubin, 2009). It is also worth noting that the adjusted method used by statistician-2 is not always correct; examples are abundant where the unadjusted method used by statistician-1 gives the correct result (Pearl, 2009b; Shrier, 2009). The correct criterion for proper choice of covariates for adjustment is given by the back-door condition (Pearl, 1993) and is the same as that deployed in the resolution of Simpson's paradox (Pearl, 2014b).

5 From Weight Gain to Birth Weight

The problem of managing differential base-rates is pervasive in all the empirical sciences. Whenever the responses of two or more groups to a treatment or a stimulus are compared, it is essential to adjust (or allow) for initial differences among those groups. The merits of adjusting for such differences were noted as far back as Fisher (1935)

> "For example, in a feeding experiment with animals, where we are concerned to measure their response to a number of different rations or diets, ... it may well be that the differences in initial weight constitute an uncontrolled cause of variation among the responses to treatment, which will sensibly diminish the precision of the comparisons" (Fisher, 1935, p. 168).
>
> "They may, however constitute an element of error which it is desirable, and possibly, to eliminate. The possibility arises from the fact that, without being equalised, these differences of initial weight may none the less be measured. Their effects upon our final results may approximately be estimated, and the results adjusted in accordance with the estimated effects, so as to afford a final precision, in many cases, almost as great as though complete equalisation had been possible" (Fisher, 1935, pp. 168–169).

In modern data analysis, the problem continued to haunt researchers across many disciplines. For example, in studying the effect of stimulus on the heart rates of rats of different ages, researchers found that the effect was different for young rats than for older rats. But their baseline heart rates were also quite different. They asked, "How are we to adjust heart-rate data obtained after an experimental treatment, for differences among animals in their base rates" (Wainer, 1991). Likewise, in studying the differential effect of schooling on white and black students, the question arises whether one should adjust for the difference of admission test scores between black and white students (Wainer and Brown, 2007). Lord himself recognized the generality of the problem as it surfaced in educational testing:

> "For example, a group of underprivileged students is to be compared with a control group on freshman grade-point average (y). The underprivileged group has a considerably lower mean grade-point average than the control group. However, the underprivileged group started college with a considerably lower mean aptitude score (x) than did the control group. Is the observed difference between the groups on y attributable to initial differences on x? Or shall we conclude that the two groups achieve differently even after allowing for initial differences in measured aptitude?" (Lord, 1969, p. 336)

Lord specifically chose x (aptitude score) and y (grade point average) to be two different variables, measured on different scales, to prevent the temptations to focus on their difference, $y - x$, as the target of interest (as statistician-1 did in the weight gain example.) In his examples, y and x can be arbitrary variables, and still, "the investigator wishes to make an "adjustment" to cancel out the effect of preexisting differences between the two groups on some other variable x" (Lord, 1969, p. 336).

Lord also raised the methodological question as to why anyone would wish "to cancel out the effect" on x. His answer was that, in certain situations we may be in possession

of practical means of suppressing the differences in x, and we wish to know if the group difference *in itself* would produce differences in y. His example was an agricultural experiment in which a given treatment shows an effect on yield (y) but also on other conditions (e.g., plant height) that can be controlled physically (e.g, by a certain fertilizer). The question then is whether the effort and expense associated with such physical control would be justified, given what we know from the data at hand. These decision-theoretic considerations have indeed been cited as the core of causal mediation analysis (Pearl, 2001, 2014b), where the value of estimating the indirect effect is tied to our ability to suppress it (or suppress the direct effect).

As mentioned earlier, the generic problem posed by Lord's paradox was initially addressed by researchers following the potential outcome framework (Holland and Rubin, 1983; Wainer, 1991; Holland, 2005; Wainer and Brown, 2007). However, lacking graphical tools for guidance, these analyses left Lord's challenge in a state of stalemate and indecision, concluding merely that the choice between the two methods of analysis depends on untestable assumptions; the problem of deciding this choice in cases where qualitative models are available remained open.

The challenge has more recently been picked up in the health sciences, where graphical tools are deployed to great advantage (Glymour, 2006; Arah, 2008; Tu et al., 2008). Here, Lord's paradox has surfaced through a variant named the Birth Weight paradox, which presents a new twist. Whereas in Lord's setup we faced a clash between two, seemingly legitimate methods of analysis, in the Birth Weight paradox we face a clash between a valid method of analysis (ANCOVA) and the scientific plausibility of its conclusion.

6 The Birth Weight Paradox

The birth-weight paradox concerns the relationship between the birth weight and mortality rate of children born to tobacco smoking mothers. It is dubbed a "paradox" because, contrary to expectations, low birth-weight children born to smoking mothers have a lower infant mortality rate than the low birth weight children of non-smokers (Wilcox, 2006).

Traditionally, low birth weight babies have a significantly higher mortality rate than others (it is in fact 100-fold higher). Research also shows that children of smoking mothers are more likely to be of low birth weight than children of non-smoking mothers. Thus, by extension the child mortality rate should be higher among children of smoking mothers. Yet real-world observation shows that low birth weight babies of smoking mothers have a lower child mortality than low birth weight babies of non-smokers.

At first sight these findings seemed to suggest that, at least for some babies, having a smoking mother might be beneficial to one's health. However, this is not necessarily the case; the paradox can be explained as an instance of "collider bias" (Cole et al., 2010) or "explain away" effect (Kim and Pearl, 1983).[6] The reasoning goes as follows: smoking may be harmful in that it contributes to low birth weight, but other causes of low birth weight

[6]Other names for this effect are "Berkson paradox," or "Berkson fallacy" (Berkson, 1946), which characterizes the general phenomenon whereby two independent causes become dependent upon observing their common effect. This phenomenon is the basis of the d-separation criterion in graphical models (Pearl, 1988, 2009a).

are generally more harmful. Now consider a low weight baby. The reason for its low weight can be either a smoking mother or those other causes. However, finding that the mother smokes "explains away" the low weight and reduces the likelihood that those "other causes" are present. This reduces the mortality rate due those other causes; smoking remains the likely cause of mortality, which is less dangerous. The net result being a lower mortality rate among low weight babies whose mother smokes, compared with with those whose mother does not smoke (Hernández-Díaz et al., 2006).

This phenomenon can easily be seen in the model of Fig. 6. We can explain it from

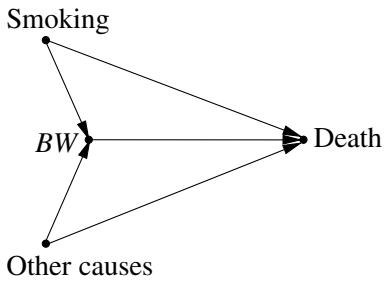

Figure 6: Showing birth weight (BW) as a "collider" affected by two independent causes: "Smoking" and "Other causes." Observing one cause (e.g., Smoking) explains away the other and reduces its probability.

two perspectives. First, we can ask for the causal effect of birth weight on death. In this context, we see that the desired effect is confounded by both Smoking and Other causes, and if we control for Smoking, it still leaves the other confounder uncontrolled, resulting in bias. Moreover, controlling for Smoking changes the probability of "Other causes" (through the collider at BW) in any stratum of BW. In particular, for underweight babies, $BW = Low$, if we compare smoking with non-smoking mothers, we would be comparing babies for which "Other causes" are rare with those for which "Other causes" are likely to occur (in order to explain the low birth weight condition.) Now, since those "Other causes" may be more dangerous to survival, we get the illusion that mortality rate increases for non-smoking mothers.

The second perspective places the birth weight example in the context of Lord's paradox and asks for the effect of smoking on mortality, discounting its effect on birth weight. Paraphrased in Lord's counterfactual language, "The researcher wants to know" how the mortality rate of babies of smoking mothers would have compared to that of non-smoking mothers, if there had been no preexisting uncontrolled differences in birth weight." Note that this question turns the problem into a mediation exercise, as in Lord's original problem (Fig. 2) and our task is to estimate the direct effect of Smoking on Death, unmediated by Birth Weight.

There is however a structural difference between the mediation model of Fig. 2 and the one in Fig. 6. Whereas in Fig. 2 we assumed no hidden confounders, such confounders are present in Fig. 6, labeled "Other causes." This makes a qualitative difference in our ability to estimate the direct effect. Adjusting for the mediator (BW) no longer severs all paths

traversing the mediators, it actually opens a new path:

$$Smoking \rightarrow \boxed{\text{BW}} \leftarrow Other\ causes \rightarrow Death,$$

by conditioning on the collider at BW. This path is spurious (i.e., non causal) and hence produces bias.

A simple way of seeing this is to recall that conditioning on the event $BW = Low$ does not physically prevent BW from changing; it merely filters out from the analysis all babies except those with $BW = Low$. Therefore, as we compare smoking with non-smoking mothers for babies of equal birth weight we are actually comparing babies with no "Other causes" to babies for whom "Other causes" are present. This of course will create an illusionary increase in mortality rates for babies of non-smoking mothers, thus explaining the Birth Weight paradox.

The fallibility of estimating direct effects by conditioning on (or "co-varying away") the mediator has been noted for quite some time (Robins and Greenland, 1992; Pearl, 1998; Cole and Hernán, 2002) and has led to modern definitions of direct and indirect effects based on counterfactual, rather than statistical conditioning (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2009). Fisher himself is reported to have failed on this question by recommending the use of ANCOVA (conditioning) to "allow" for variations in the mediator (Fisher, 1935, p. 165; Rubin, 2005). Fisher's blunder led Rubin to conclude that "the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful to clear statistical thinking" (Rubin, 2004). As a result, Frangakis and Rubin (2002) proposed alternative definitions of direct and indirect effects based on "principal strata" which, ironically, suffer from at least as many problems as Fisher's (Pearl, 2011; VanderWeele, 2011).

The Birth Weight paradox was instrumental in bringing this controversy to a resolution. First, it has persuaded most epidemiologists that collider bias is a real phenomenon that needs to be reckoned with (Cole et al., 2010). Second, it drove researchers to abandon traditional mediation analysis (usually connected with Judd and Kenny (1981a) and Baron and Kenny (1986)) in which mediation is define by statistical conditioning (or "statistical control," in which the mediator is "partialled out"), and replace it with causally defined mediation analysis based on counterfactual conditioning (VanderWeele, 2009; Imai et al., 2010; Pearl, 2012; Valeri and VanderWeele, 2013; Pearl, 2014a; Muthén, 2014). I believe Frederic Lord would be mighty satisfied today with the development that his 1967 observation has spawned.

Acknowledgment

This paper benefitted from discussions with Ian Shrier, Howard Wainer, Steven Cole, and Felix Theome. This research was supported in parts by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665 and #N00014-10-1-0933.

References

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>.

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

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

Bock, R. (1975). *Multivariate statistical methods in behavioral research*. McGraw-Hill, NY.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31** 163–165.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D. and Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* **39** 417–420.

Cox, D. and McCullagh, P. (1982). A biometrics invited paper with discussion. some aspects of analysis of covariance. *Biometrics* **38** 541–561.

Eriksson, K. and Häggström, O. (2014). Lord's paradox in a continuous setting and a regression artifact in numerical cognition research. *PLOS One* **9** artikel nr e95949.

Fisher, R. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics* **1** 21–29.

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In *Methods in Social Epidemiology*. John Wiley and Sons, San Francisco, CA, 393–428.

Hernández-Díaz, S., Schisterman, E. and Hernán, M. (2006). The birth weight "paradox" uncovered? *American Journal of Epidemiology* **164** 1115–1120.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81** 945–960.

Holland, P. and Rubin, D. (1983). On Lord's paradox. In *Principals of Modern Psychological Measurement* (H. Wainer and S. Messick, eds.). Lawrence Earlbaum, Hillsdale, NJ, 3–25.

Holland, P. W. (2005). Lord's paradox. In *Encyclopedia of Statistics in Behavioral Science* (B. S. Everitt and D. Howell, eds.). Wiley, New York, 1106–1108.

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

Judd, C. and Kenny, D. (1981a). *Estimating the Effects of Social Interactions*. Cambridge University Press, Cambridge, England.

Judd, C. and Kenny, D. (1981b). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5** 602–619.

Kim, J. and Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*. Karlsruhe, Germany.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin* **68** 304–305.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin* **72** 336–337.

Muthén, B. (2014). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Tech. rep., Graduate School of Education and Information Studies, University of California, Los Angeles, CA. Forthcoming, *Psychological Methods*.

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. (1994). A probabilistic calculus of actions. In *Uncertainty in Artificial Intelligence 10* (R. L. de Mantaras and D. Poole, eds.). Morgan Kaufmann, San Mateo, CA, 454–462.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research* **27** 226–284.

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

Pearl, J. (2009b). Remarks on the method of propensity scores. *Statistics in Medicine* **28** 1415–1416. <http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf>.

Pearl, J. (2011). Principal stratification a goal or a tool? *The International Journal of Biostatistics* **7**. Article 20, DOI: 10.2202/1557-4679.1322. Available at: http://www.bepress.com/ijb/vol7/iss1/20.

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science* **13** 426–436. <http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf>.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* **1** 155–170.

Pearl, J. (2014a). Interpretation and identification of causal mediation. *Psychological Methods* **19** 459–481.

Pearl, J. (2014b). Understanding Simpson's paradox. *The American Statistician* **88** 8–13.

Pearl, J. (2016). The sure-thing principle. *Journal of Causal Inference,* Causal, Casual, and Curious Section **4** 81–86.

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

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine* **28** 1420–1423.

Savage, L. (1962). *The Foundations of Statistical Inference: A Discussion.* John Wiley and Sons, Inc., New York, NY.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine* **25** 4334–4344.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* **9** 1941–1979.

Shrier, I. (2009). Letter to the editor: Propensity scores. *Statistics in Medicine* **28** 1317–1318. See also Pearl 2009 <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

Tu, Y.-K., D.J. G. and Gilthorpe, M. (2008). Simpson's paradox, Lord's paradox, and suppression effects are the same phenomenon – the reversal paradox. *Emerging Themes in Epidemiology* **5** 2.

Valeri, L. and VanderWeele, T. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods* **13** Epub ahead of print.

van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research* **48** 895–922.

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

VanderWeele, T. J. (2011). Principal stratification – uses and limitations. *The International Journal of Biostatistics* **7** 1–14.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin* **109** 147–151.

Wainer, H. and Brown, L. M. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. In *Handbook of Statistics 26: Psychometrics* (C. Rao and S. Sinharay, eds.), vol. 26. Elsevier B.V., North Holland, 893–918.

Wilcox, A. (2006). The perils of birth weight – a lesson from directed acyclic graphs. *American Journal of Epidemiology* **164** 1121—1123.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.