The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Two Statistical Paradoxes in the Interpretation of Group Differences

Howard Wainer & Lisa M Brown

To cite this article: Howard Wainer & Lisa M Brown (2004) Two Statistical Paradoxes in the Interpretation of Group Differences, The American Statistician, 58:2, 117-123, DOI: 10.1198/0003130043268

To link to this article: http://dx.doi.org/10.1198/0003130043268

Published online: 01 Jan 2012.

Two Statistical Paradoxes in the Interpretation of Group Differences: Illustrated with Medical School Admission and Licensing Data

Howard WAINER and Lisa M. BROWN

Interpreting group differences observed in aggregated data is a practice that must be done with enormous care. Often the truth underlying such data is quite different than a naïve first look would indicate. The confusions that can arise are so perplexing that some of the more frequently occurring ones have been dubbed paradoxes. This article describes two of these paradoxes—Simpson's paradox and Lord's paradox—and illustrates them in a single dataset. The dataset contains the score distributions, separated by race, on the biological sciences component of the Medical College Admission Test (MCAT) and Step 1 of the United States Medical Licensing Examination™ (USMLE). Our goal in examining these data was to move toward a greater understanding of race differences in admissions policies in medical schools. As we demonstrate, the path toward this goal is hindered by differences in the score distributions which gives rise to these two paradoxes. The ease with which we were able to illustrate both of these paradoxes within a single dataset is indicative of how widespread they are likely to be in practice.

KEY WORDS: Group differences; Lord's paradox; Medical College Admission Test; Rubin's model for causal inference; Simpson's paradox; Standardization; United States Medical Licensing Examination.

Howard Wainer is Distinguished Research Scientist, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104 (E-mail: hwainer@nbme.org). Lisa M. Brown is Psychometrician, American Board of Internal Medicine. We are grateful to the Association of American Medical Colleges for allowing us to use their MCAT data. More specifically we thank Ellen Julian for her help in providing us with the information we required. In addition, this entire project was prompted by Don Melnick, who not only suggested that we do it but who has supported our inquiry. We are not unaware of the sensitive nature of some of the questions we are asking and fully appreciate Don's support and helpful comments on an earlier draft of this article. David Swanson and Douglas Ripkey generously provided us with a file of the USMLE results paired with MCAT scores; obviously a key element in this investigation. An earlier version of this article was read and commented on by our colleagues Ron Nungester and Brian Clauser; we thank them for helping us excise errors and fuzzy thinking. We are also grateful for the insightful comments of James Albert and his wise, but anonymous, associates. A longer version of this article is available online at www.Statlit.org/PDF/2004Wainer_ThreeParadoxes.pdf This research was paid for by NBME and we thank our employer for providing such fascinating opportunities. Of course, the opinions expressed here are ours, as are any errors we may have made along the way. This work is collaborative in every respect and the order of authorship is random. Last, our gratitude to Paul Holland and Don Rubin for explaining Lord's Paradox to us, and for their permission to use their explanation in Section 4.

To count is modern practice, the ancient method was to guess
—Samuel Johnson

Evidence may not buy happiness, but it sure does steady the nerves
—paraphrasing Satchel Paige's comment about money

1. INTRODUCTION

Modern policy decisions involving group differences are both based on, and evaluated by, empirical evidence. But the understanding and interpretation of the data that comprise such evidence must be done carefully, for many traps await the unwary. This essay explores two statistical paradoxes that can potentially mislead us and illustrates these paradoxes with data used in the admission of candidates to medical school, and one measure of the success of those admissions.

The first is known as Simpson's paradox (Yule 1903) and appears when we look at the aggregate medical school application rates by ethnicity. The second paradox, which was first described by Lord (1967), emerges when we try to estimate the size of the effect of medical school training on students.

The balance of this essay is laid out as follows. Section 2 describes the data that form the basis of our investigation and provides some summarizations. Section 3 describes Simpson's paradox and demonstrates its existence within our data and shows how to ameliorate its effects through the method of standardization. Section 4 demonstrates Lord's paradox and describes how its puzzling result can be understood by embedding the analysis within Rubin's model for causal inference. Section 5 concludes with a discussion of these findings.

2. THE DATA

There are many steps on the path toward becoming a physician. Two important ones that occur early on are tests. The first test, the Medical College Admission Test (MCAT), is usually taken during the junior or senior year of college and is one important element in gaining admission to medical school. The second test is Step 1 of the United States Medical Licensing Exam (USMLE). Step 1 is the first of a three-part exam a physician must pass to become licensed in the United States. This test is usually taken after the second year of medical school and measures the extent to which an examinee understands and can apply important concepts of the basic biomedical sciences. For the purposes of this investigation we examined the performance of all black and white examinees whose most recent MCAT was taken during the three-year period between 1993 and 1995. Two samples of examinees tested during this time were used in the

 DOI: 10.1198/0003130043268

Table 1. Selected Medical School Application and Licensing Statistics

	Frequencies											
	Last MCAT score all MCAT takers 1993–1995			Applied to medical school 1994–1996			Accepted at medical school 1994–1996			USMLE Step 1 test volumes 1996–1998		
MCAT-BS score	Black	White	Total	Black	White	Total	Black	White	Total	Black	White	Total
3 or less	1,308	1,168	2,476	404	238	642	8	1	9	6	1	7
4	1,215	2,094	3,309	482	557	1,039	52	10	62	39	10	49
5	1,219	3,547	4,766	582	1,114	1,696	202	45	247	116	36	152
6	1,269	5,289	6,558	752	1,983	2,735	417	163	580	256	140	396
7	1,091	6,969	8,060	748	3,316	4,064	518	636	1,154	338	589	927
8	1,234	11,949	13,183	868	6,698	7,566	705	2,284	2,989	537	2,167	2,704
9	702	13,445	14,147	544	8,628	9,172	476	4,253	4,729	340	4,003	4,343
10 or more	660	29,752	30,412	511	20,485	20,996	475	14,244	14,719	334	12,786	13,120
Totals	**8,698**	**74,213**	**82,911**	**4,891**	**43,019**	**47,910**	**2,853**	**21,636**	**24,489**	**1,966**	**19,732**	**21,698**

	Selected conditional probabilities											
	Probability of MCAT taker applying to medical school			Probability of MCAT taker being accepted to medical school			Probability of MCAT taker taking USMLE Step 1			Probability of medical school acceptee taking USMLE Step 1		
MCAT-BS score	Black	White	Total	Black	White	Total	Black	White	Total	Black	White	Total
3 or less	0.31	0.20	0.26	0.01	0.00	0.00	0.00	0.00	0.00	0.75		0.78
4	0.40	0.27	0.31	0.04	0.00	0.02	0.03	0.00	0.01	0.75	1.00	0.79
5	0.48	0.31	0.36	0.17	0.01	0.05	0.10	0.01	0.03	0.57	0.80	0.62
6	0.59	0.37	0.42	0.33	0.03	0.09	0.20	0.03	0.06	0.61	0.86	0.68
7	0.69	0.48	0.50	0.47	0.09	0.14	0.31	0.08	0.12	0.65	0.93	0.80
8	0.70	0.56	0.57	0.57	0.19	0.23	0.44	0.18	0.21	0.76	0.95	0.90
9	0.77	0.64	0.65	0.68	0.32	0.33	0.48	0.30	0.31	0.71	0.94	0.92
10 or more	0.77	0.69	0.69	0.72	0.48	0.48	0.51	0.43	0.43	0.70	0.90	0.89
Totals	**0.56**	**0.58**	**0.58**	**0.33**	**0.29**	**0.30**	**0.23**	**0.27**	**0.26**	**0.69**	**0.91**	**0.89**

analyses. The first sample of approximately 83,000 scores comprises all black and white examinees whose most recent MCAT was taken during this time. This sample includes all examinees rather than being limited to only those applying to medical school. Additionally, because the sample reflects performance of examinees who had taken the MCAT after repeated attempts, the initial scores from low scoring examinees who repeated the examination to improve their performance were not included. This makes these average scores somewhat higher than those reported elsewhere (http://www.aamc.org).

The funnel of medical school matriculation continued with about 48,000 (58%) of those who took the MCAT actually applying to medical school; of these about 24,000 (51%) were actually accepted. And finally, approximately 22,000 (89%) of the candidates who were accepted to allopathic medical schools, sat for Step 1 three years after their last MCAT attempt. By limiting our sample to those who entered medical school the year after taking the MCAT and took Step 1 two years later, we have excluded those who progressed through these steps in less typical amounts of time. But this seems like a plausible way to begin, and the conclusions we reach using this assumption should not be very far from the truth.

Table 1 presents the distributions of MCAT-Biological Sciences scores for two racial groups along with selected conditional probabilities. (MCAT is a test that consists of four parts—Verbal Reasoning, Physical Sciences, Biological Sciences, and a Writing Sample. The Biological Sciences score is the one that correlates most highly with subsequent performance on Step 1 of the USMLE, and so we used it as the stratifying variable throughout our study. None of our conclusions would be changed if we used an amalgam of all parts of the test, but the interpretations could get more complex. Therefore henceforth when we use the term "MCAT" we mean "MCAT Biological Sciences.") The first column in the upper portion of Table 1 shows the MCAT scores; we grouped some adjacent extreme score categories together because the sample sizes in the separate categories were too small in one or the other of the two groups to allow reliable inferences. The first section of the table shows the distributions of MCAT scores by race for black and white candidates whose most recent attempt was between 1993 and 1995. The second and third sections present the number of examinees from each group who applied to allopathic medical schools the following year and the respective acceptance rates. The final section shows the distribution of MCAT scores among those in our sample who matriculated to medical school and took Step 1 of the USMLE three years after their last MCAT attempt.

The bottom portion of Table 1 presents selected conditional probabilities at each level of MCAT score that were derived from the frequencies in the top portion in the indicated fashion.

For the purposes of this discussion there are three important characteristics of Table 1: (1) the higher the MCAT score the greater the likelihood of applying to medical school, being selected, and eventually taking Step 1; (2) at every MCAT score level the proportion of black MCAT takers taking Step 1 is higher than for white applicants; and (3) despite this, the Step 1 rates for whites overall was higher than for blacks. If we have not made any errors in our calculations, how do we account for this remarkable result? Are black students sitting for the licensing

Table 2. NAEP 1992 8th Grade Math Scores

	State	White	Black	Other Nonwhite	Standardized
Nebraska	**277**	281	236	259	**271**
New Jersey	**271**	283	242	260	**273**
		Proportion of population			
Nebraska		87%	5%	8%	
New Jersey		66%	15%	19%	
Nation		69%	16%	15%	

exam with greater likelihood than whites? Or with lesser? This is an example of Simpson's paradox and in the next section we discuss how it occurs and show how we can ameliorate its effects.

3. SIMPSON'S PARADOX

The seeming anomaly in Table 1 is not rare. It shows up frequently when data are aggregated and is well known among statisticians (e.g., Blyth 1972; Rinott and Tamm 2003; Samuels 1993; Simpson 1951; Wainer 1986; Westbrooke 1998). Indeed we see it also in the probabilities of applying to medical school. Let us examine another, simpler, example to help us understand both how it occurs and what we can do to allow us to make sensible inferences from such results.

As our second example, consider the results from the National Assessment of Educational Progress shown in Table 2. We see that 8th grade students in Nebraska scored six points higher in mathematics than their counterparts in New Jersey. Yet we also see that both white and black students do better in New Jersey. Indeed, all other students do better in New Jersey as well. How is this possible? Once again it is an example of Simpson's paradox. Because a much greater percentage of Nebraska's 8th grade students (87%) are from the higher scoring white population than in New Jersey (66%), their scores contribute more to the total.

Given these results, we could ask, "Is ranking states on such an overall score sensible?" It depends on the question that these scores are being used to answer. If the question is something like "I want to open a business. In which state will I find a higher proportion of high-scoring math students to hire?" this unadjusted score is sensible. If, however, the question of interest is "I want to enroll my children in school. In which state are they likely to do better in math?" a different answer is required. If your children have a race (it does not matter what race), they are likely to do better in New Jersey. If questions of this latter type are the ones that are asked more frequently, it makes sense to adjust the total to reflect the correct answer. One way to do this is through the method of standardization, in which we calculate what each state's score would be if it were based upon a common demographic mixture. In this instance one sensible mixture to use is that of the nation overall. Thus, after standardization the result obtained is the score we would expect each state to have if it had the same demographic mix as the nation. To create the standardized score for New Jersey we multiply the average score for each subgroup by their respective percentages in the nation, for example, $(283 \times 0.69) + (242 \times 0.16) + (260 \times 0.15) = 273$. Because New Jersey's demographic mix is not very different

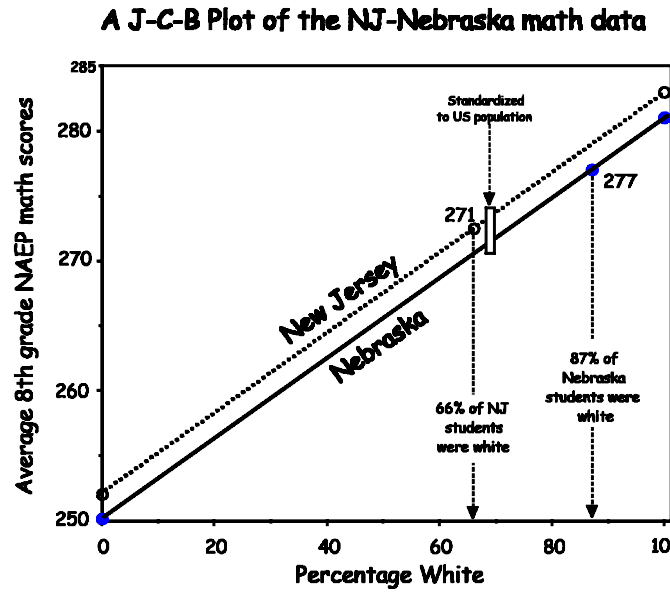

Figure 1. A graph developed by Jeon, Chung, and Bae (1987) that illuminates the conditions for Simpson's paradox as well as how standardization ameliorates it.

from the national mix, its score is not affected much (273 instead of 271), whereas because of Nebraska's largely white population its score shrinks substantially (271 instead of 277).

Simpson's paradox is illuminated through a clever graphic developed by Jeon, Chung, and Bae (1987) (and independently reinvented by Baker and Kramer 2001). In Figure 1 the solid line represents what Nebraska's average score would be with any proportion of white students. The solid point at "87% white" shows what the score was with the actual percentage. Similarly, the dashed line shows what New Jersey's average score would be for any percentage of whites, with the unshaded point showing the actual percentage. We can readily see how Nebraska's average point is higher than New Jersey's. The unshaded rectangle represents what both states' averages would be with a hypothetical population of 69% white—the standardization mix. This plot shows that what particular mixture is chosen for standardization is irrelevant to the two state's relative positions, because the two states' lines are parallel.

The use of standardization is not limited to comparing different states with one another. Indeed it may be even more useful comparing a state with itself over time. If there is a change in educational policy (e.g., per pupil expenditure) standardization to the demographic structure of the state at some fixed point in time allows us to estimate the effect of the policy change uncontaminated by demographic shifts.

Now we can return to the data about MCAT examinees in Table 1 with greater understanding. Why is it that the overall rate for taking Step 1 is lower for blacks than for white examinees, when we see that the rate is higher for blacks (often markedly higher) at each MCAT score level? The overall rate of 23% for black students is caused by a combination of two factors: policy and performance. For many policy purposes it would be well if we could disentangle these effects. As demonstrated in the prior example, one path toward clarity lies in standardization. If we wish to compare the rates for black and white students

Table 3. Distributions of Step 1 Rates by Ethnicity with Standardized Totals

	Step 1 rates		Percentage of		Standardized Step 1 rates	
MCAT score	White	Black	Whites	Blacks	Whites	Blacks
3 or less	0.1%	0.5%	**1.6%**	15.0%	0.0%	0.0%
4	0.5%	3.2%	**2.8%**	14.0%	0.0%	0.1%
5	1.0%	9.5%	**4.8%**	14.0%	0.0%	0.5%
6	2.6%	20.2%	**7.1%**	14.6%	0.2%	1.4%
7	8.5%	31.0%	**9.4%**	12.5%	0.8%	2.9%
8	18.1%	43.5%	**16.1%**	14.2%	2.9%	7.0%
9	29.8%	48.4%	**18.1%**	8.1%	5.4%	8.8%
10 or more	43.0%	50.6%	**40.1%**	7.6%	17.2%	20.3%
Total	26.6%	22.6%			26.6%	41.0%

that current policy generates we must rid the summary of the effects of differential performance and estimate standardized rates. Standardized rates can be obtained by multiplying the Step 1 rates of each stratum of both black and white students by the score distribution of white students. Multiplying the two columns of Step 1 rates in Table 3 by the score distribution of whites (in bold) yields the two final columns, which when summed are the standardized rates; standardized to the white score distribution. Of course the white summary stays the same 27%, but the standardized Step 1 rate for black students is 41%. We can use this information to answer the question:

If black students scored the same on the MCAT as white students what proportion would go on to take Step 1?

Comparing the total Step 1 rates for blacks and whites after standardization reveals that if black and white candidates performed equally well on the MCAT, blacks would take Step 1 at a rate 54% higher than whites. The standardization process also allows for another comparison of interest. The difference between the standardized rate of 41% for blacks and the actual rate of 23% provides us with the effect of MCAT performance on Step 1 rates of black students. This occurs because white students are more heavily concentrated at high MCAT scores, which have a higher rate of taking Step 1. Standardization tells us that if black students had that same MCAT distribution their rate of taking Step 1 would almost double.

Standardization is not the only way to make sense of data that displays Simpson's paradox. If we do not need to provide a single summary statistic for each group, it is often sensible to just report the conditional means. Indeed this is a sensible approach when one of the entities being compared is missing one or more of the component groups. For example, if we substituted Montana for Nebraska in the second example we could not standardize because of the almost nonexistent size of non-Hispanic Blacks in Montana. But we could compare the two states one-group-at-a-time without misleading anyone.

4. LORD'S PARADOX

We have observed that the performance of the two groups on the outcome variable, the USMLE Step 1 score, depends on both performance on the predictor variable, the MCAT score, and on group membership. Faced with this observation it is natural to ask:

How much does group membership matter in measuring the effect of medical school?

What does this question mean? One plausible interpretation would be to examine an individual's rank among an incoming class of medical students, and then examine her rank after receiving a major portion of her medical education. If her rank did not change, we could conclude that the effect of medical school was the same for that individual as it was for the typical medical student. If we wish to measure the effect of medical school on any group, we might compare the average change in ranks for that group with another. But this is not the only plausible approach. Alternatively we might use the pre-medical school ranks as a covariate and examine the differences between the groups' average medical school rank after adjusting for the pre-medical school rank. How we might do this and how we interpret the results is the subject of this section. (We use ranks rather than test scores to circumvent the problems generated by the two different tests being scored on different scales and having very different reliabilities. It is not that such an alternative path could not be taken, but we felt that for this illustration it would be cleaner, simpler, and more robust to assumptions if we stuck with analyses based on the order statistics.)

We begin the investigation by:

a. Drawing a random sample from the USMLE Step 1 takers of 200 white examinees and 200 black examinees.
b. Then we rank these 400 examinees on both their MCAT scores and their Step 1 scores.
c. Next we subtract each examinee's rank on the Step 1 from that person's rank on the MCAT.
d. Then we calculate the average difference for white and for black examinees.

We found that white examinees' ranks improved, on average, about 19 places. This was, of course, balanced by a decline of 19 places in rank among black examinees, or a total differential effect of 38.

But, as we mentioned before, taking the difference in ranks is not the only way to estimate this effect. Alternatively, we could use the MCAT rank as a covariate and look at the ranks of the individuals on the adjusted USMLE Step 1 (the residuals on Step 1 ranks after a linear adjustment for MCAT score). When we did exactly this we found that white examinees' Step 1 ranks, after adjusting for MCAT scores, improved, on average, about 9 places, with black examinees' ranks declining the same 9 places, for a total differential effect of 18.

The results of these two analyses were substantially different. Which is the right answer? This question was posed previously by Fred Lord (1967) in a two-page article that clearly laid out what has since become known as Lord's paradox. He did not explain it. The problem appears to be that the analysis of covariance cannot be relied upon to properly adjust for uncontrolled preexisting differences between naturally occurring groups. A full explanation of the paradox first appeared fully 16 years later (Holland and Rubin 1983) and relies heavily on Rubin's model for causal inference (Rubin 1974).

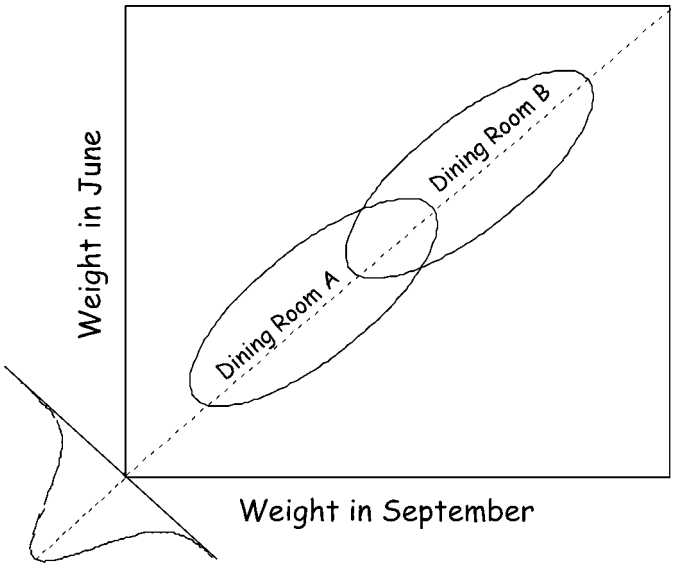

Figure 2. A graphical depiction of Lord's paradox showing the bivariate distribution of weights in two dining rooms at the beginning and end of each year augmented by the 45° line (the principal axis). The ovals represent the bivariate scatter within each dining room.

The paradox, as Lord described it, was based on the following hypothetical situation:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls .... 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. (p. 304)

Lord framed his paradox in terms of the analyses of two hypothetical statisticians who come to quite different conclusions from the data in this example.

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 Figure 2 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 Dining Room A. This result is depicted graphically in Figure 3. 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. From this result the second 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 statisticians 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?" But causal questions are always comparative (The comedian Henny Youngman's signature joke about causal inference grew from his reply to "How's your wife?" He would then quip, "Compared to what?") and the decision of how to estimate the standard of comparison is what differentiates Lord's two statisticians. Each statistician made an untestable assumption about the subjunctive situation of what would have been a student's weight in June had that student not been in the dining room of interest. This devolves directly from the notion of a causal effect being the difference between what happened under the treatment condition versus what happened under the control condition.

The fundamental difficulty with causal inference is that we can never observe both situations. Thus, we must make some sort of assumption about what would have happened had the person been in the other group. In practice we get hints of what such a number would be through averaging and random assignment. This allows us to safely assume that, on average, the experimental and control groups are the same.

In Lord's setup the explication is reasonably complex. 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.

How does this approach help us to untangle the conflicting estimates for the relative value of medical school for the two racial groups? (*Note:* This section borrows heavily from Holland and Rubin (1983, p. 5–8) and uses their words as well as their

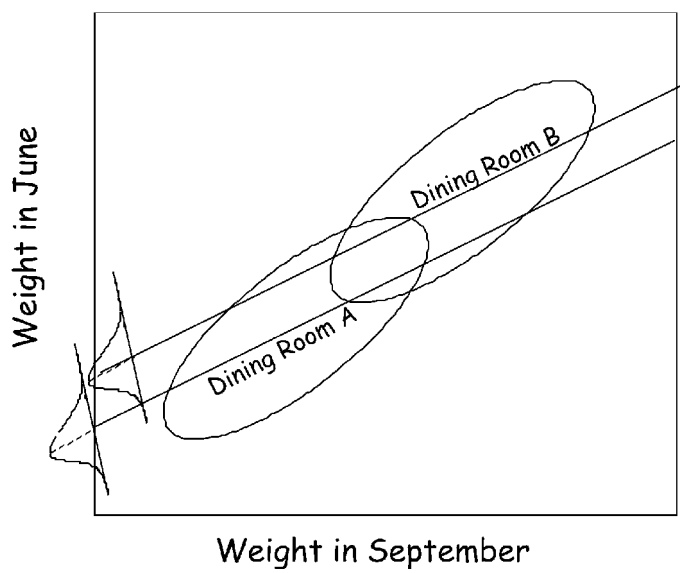

Figure 3. A graphical depiction of the second part of Lord's paradox: the result obtained by covarying out each student's September weight. The ovals represent the bivariate scatter within each dining room, the lines drawn in are the within-dining hall regression lines, and the distributions drawn in represent the distribution of residuals of June weight after adjusting for September's.

ideas.) To do this requires a little notation and some algebra. The elements of the model are:

1. a population of units, P;
2. an "experimental manipulation," with levels T and C and its associated indicator variable, S;
3. a subpopulation indicator, G;
4. an outcome variable, Y; and
5. a concomitant variable, X.

The purpose of the model is to allow an explicit description of the quantities that arise in three types of studies:

a. descriptive studies,
b. uncontrolled causal studies, and
c. controlled causal studies.

A *descriptive study* has no experimental manipulation so there is only one version of Y and X and no treatment indicator variable S.

Controlled and *uncontrolled causal studies* both have experimental manipulations and differ only in the degree of control that the experimenter has over the treatment indicator, S. In a controlled causal study, the values of S are determined by the experimenter and can depend on numerous aspects of each unit (e.g., subpopulation membership, values of covariates) but not on the value of Y, because that is observed after the values of S are determined by the experimenter. In an uncontrolled causal study the values of S are determined by factors that are beyond the experimenter's control. Critical here is the fact that in a controlled study S can be made to be statistically independent of Y_C and Y_T whereas in an uncontrolled causal study this is not true.

The causal effect of T on Y (relative to C) for each unit in P is given by the difference $Y_T - Y_C$. The average causal effect of T versus C on Y in P is $E(Y_T - Y_C)$, which equals $E(Y_T) - E(Y_C)$. This shows us how the unconditional means of Y_T and Y_C over P have direct causal interpretations. But because T and C are usually not observable on the same unit, $E(Y_T)$ and $E(Y_C)$ are not typically observable.

In a causal study, the value of Y that is observed on each unit is Y_S, so that when $S = T$, Y_T is observed and when $S = C$, Y_C is observed. The expected value of Y for the "treatment group" is $E(Y_T|S = T)$ and for the "control group" is $E(Y_C|S = C)$. There is no reason to believe that $E(Y_T)$ should equal $E(Y_T|S = T)$, or that $E(Y_C)$ should equal $E(Y_C|S = C)$. Hence neither $E(Y_T|S = T)$ nor $E(Y_C|S = C)$ have direct causal interpretation.

Consider that $E(Y_T|S = T)$ and $E(Y_T)$ are related through

$$E(Y_T) = E(Y_T|S = T)P(S = T) + E(Y_T|S = C)P(S = C). \quad (1)$$

There is the obvious parallel version connecting $E(Y_C|S = T)$ with $E(Y_C)$. The second term of (1) is not observable. This makes explicit the basis of our earlier assertion about the shortcomings of $E(Y_T|S = T)$ and $E(Y_C|S = C)$ for making direct causal interpretations.

Note that Equation (1) involves the average value of Y_T, among those units exposed to C. But $E(Y_T|S = C)$ and its parallel $E(Y_C|S = T)$ can never be directly measured except when Y_T and Y_C can both be observed on all units. This is what Holland and Rubin (1983, p. 9) termed "the fundamental problem of causal inference."

With this model laid out, let us return to the problem of measuring the differential effect of medical school.

Study Design

P: 400 medical students in the years specified
T: Went to medical school
C: Unknown

Variables Measured

G: Student Race ($W = 1$, $B = 2$)
X: The rank of a student on the MCAT
Y: The rank of a student on Step 1 of the USMLE

This layout makes clear that the control condition was undefined—no one was exposed to C ($S = T$ for all students)—and so any causal analysis must make untestable assumptions. As is perhaps obvious now, the two different answers we got to the same question must have meant that we made two different untestable assumptions. This will become visible by making the inference explicit.

The causal effect of medical school for black and white students is

$$D_i = E(Y_T - Y_C|G = i) \quad i = 1, 2, \quad (2)$$

and so the difference of average causal effects is

$$D = D_1 - D_2. \quad (3)$$

This can be expressed in terms of individual subpopulation averages,

$$D = [E(Y_T|G = 1) - E(Y_C|G = 1)] - [E(Y_T|G = 2) - E(Y_C|G = 2)]. \quad (4)$$

We can profitably rearrange this to separate the observed Y_T from the unobserved Y_C

$$D = [E(Y_T|G = 1) - E(Y_T|G = 2)] - [E(Y_C|G = 1) - E(Y_C|G = 2)]. \quad (5)$$

The first approach estimated the effect of medical school by just looking at the difference in the ranks on MCAT and Step 1. Doing so made the (entirely untestable) assumption that an individual's response to the control condition, whatever that might be, is given by his/her rank on the MCAT

$$Y_C = X \quad (6)$$

yielding,

$$E(Y_C|G = i) = E(X|G = i). \quad (7)$$

The second approach estimated the effect of medical school by using the students' rank on the MCAT as a covariance adjustment, which corresponds to the following two conditional expectations:

$$E(Y_T|X, G = i) \quad i = 1, 2, \quad (8)$$

and the mean, conditional, improvement in rank in group i at X is

$$U_i(X) = E(Y_T - X|X, G = i) \quad i = 1, 2. \tag{9}$$

Hence, the difference in these conditional ranks at X is

$$U(X) = U_1(X) - U_2(X). \tag{10}$$

The second analysis assumes that the conditional expectations in (8) are linear and parallel. Thus, we can write

$$E(Y_T|X, G = i) = a_i + bX \quad i = 1, 2 \tag{11}$$

Substituting into (9) yields

$$U_i(X) = a_i + (b - 1)X \quad i = 1, 2. \tag{12}$$

And hence (10) simplifies to

$$U(X) = a_1 - a_2. \tag{13}$$

The second approach correctly interprets $U(X)$ as the average amount that a white student's ($G = 1$) rank will improve over a black student ($G = 2$) of equal MCAT score. This is descriptively correct, but has no direct causal interpretation since U is not directly related to D. To make such a connection we need to make the untestable assumption, related to (6) that

$$Y_C = a + bX. \tag{14}$$

Where b is the common slope of the two within-groups regression lines in (11). This allows the interpretation of $U(X)$ as the difference in the causal effects D in Equation (3).

Both of these assumptions seem to stretch the bounds of credulity, but (14) seems marginally more plausible. However deciding this issue was not our goal. Instead we wished to show how subtle an argument is required to unravel this last paradox in the investigation of group differences. The interested reader is referred to Holland and Rubin (1983) or Wainer (1991) for a fuller description of how Rubin's model of causal inferences helps us to understand this subtle paradox.

5. CONCLUSION

"What we don't know won't hurt us,
it's what we do know that ain't"

—Will Rogers

This essay, and the research behind it, has two goals. The first is to publicize more broadly the pitfalls that await those who try to draw inferences from observed group differences. The second is to provide analytic tools to allow the construction of bridges over those pitfalls.

Group differences must be examined if we wish to evaluate empirically the efficacy of modifications in policy. But such comparisons, made naïvely, are very likely to lead us astray.

Ridding ourselves of Simpson's paradox by disaggregation, or through the use of standardization is straightforward. But we must always remember that there may be another, unnoticed, variable that could reverse things again. Inferences must be made carefully. The only reasonably certain way to be sure that stratification by some unknown variable will not reverse your inference is to have random assignment to groups. When assignment is not random the possibility of Simpson's paradox is always lurking in the background.†

Lord's paradox is the latest of this pair. It occurs when data analysts use their favorite method to assess group differences without careful thought about the question they are asking. It is, by far, the most difficult paradox to disentangle and requires clear thinking. It also emphasizes how the assessment of group differences often entails making untestable assumptions. This too should give us pause when we try to draw strong conclusions.

† Benjamin Disraeli (1804–1881) was twice prime minister of England (1868, 1874–1880). At an earlier time in his career he was an outspoken critic of Sir Robert Peel's (1788–1850) free-trade policies, and to support his criticism he offered data defending the Corn Laws (1845). Peel offered counter data that justified his desire to repeal them. The two sets of data seemed contradictory, and, it is said that Disraeli, not knowing about Simpson's Paradox (or the use of standardization to correct it), exclaimed out of frustration, "Sir, there are lies, damn lies, and statistics."

[Received April 2003. Revised February 2004.]

REFERENCES

Baker, S. G., and Kramer, B. S. (2001), "Good for Women, Good for Men, Bad for People: Simpson's Paradox and the Importance of Sex-Specific Analysis in Observational Studies," *Journal of Women's Health and Gender-Based Medicine*, 10, 867–872.

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

Holland, P. W., and Rubin, D. B. (1983), "On Lord's Paradox," in *Principals of Modern Psychological Measurement*, eds. H. Wainer and S. Messick, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 3–25.

Jeon, J. W., Chung, H. Y., and Bae, J. S. (1987), "Chances of Simpson's Paradox," *Journal of the Korean Statistical Society*, 16, 117–125.

Lord, F. M. (1967), "A Paradox in the Interpretation of Group Comparisons," *Psychological Bulletin*, 68, 304–305.

Rinott, Y., and Tam, M. (2003), "Monotone Regrouping, Regression, and Simpson's Paradox," *The American Statistician*, 57, 139–141.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Samuels, M. L. (1993), "Simpson's Paradox and Related Phenomena," *Journal of the American Statistical Association*, 88, 81–88.

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

Wainer, H. (1986) "Minority Contributions to the SAT Score Turnaround: An Example of Simpson's Paradox," *Journal of Educational Statistics*, 11, 239–244.

——— (1991), "Adjusting for Differential Base-Rates: Lord's Paradox Again," *Psychological Bulletin*, 109, 147–151.

Westbrooke, I. (1998), "Simpson's Paradox: An Example in a New Zealand Survey of Jury Composition," *Chance*, 11, 40–42.

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