Handbook of Statistics, Vol. 26
ISSN: 0169-7161
2007 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(06)26028-0

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

Howard Wainer and Lisa M. Brown

Abstract

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. In this chapter we describe three of the best known of these paradoxes – Simpson's Paradox, Kelley's Paradox, and Lord's Paradox – and illustrate them in a single data set. The data set 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 three paradoxes. The ease with which we were able to illustrate all of these paradoxes within a single data set is indicative of how wide spread they are likely to be in practice.

"*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. In this

chapter we explore three statistical paradoxes that can potentially mislead us and illustrate 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; Simpson, 1951) and appears when we look at the aggregate medical school application rates by ethnicity. The second is Kelley's Paradox (Wainer, 2000), which shows its subtle effect when we examine the success rates of minority medical students on Step 1 of the U.S. Medical Licensing Exam (USMLE-1). And, finally, the third paradox, which was first described by Lord (1967), emerges when we try to estimate the effect size of medical school training on students.

The balance of this chapter is laid out as follows: in Section 2 we describe the data that form the basis of our investigation and provide some summarizations; in Section 3 we describe Simpson's Paradox and demonstrate its existence within our data and show how to ameliorate its effects through the method of standardization; in Section 4 we describe Kelley's Paradox and use our data to illustrate its existence; in Section 5 we demonstrate Lord's paradox and describe how its puzzling result can be understood by embedding the analysis within Rubin's model for causal inference. We conclude in Section 6 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 who's most recent MCAT was taken during the three-year period between 1993 and 1995. Two samples of examinees testing during this time were used in the 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, our second sample of 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.

In Table 1 we present the distributions of MCAT-Biological Sciences[1] scores for two racial groups along with selected conditional probabilities. 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 on the top portion in the indicated fashion.

For the purposes of this discussion there are three important characteristics of Table 1: (i) the higher the MCAT score the greater the likelihood of applying to medical school, being selected, and eventually taking Step 1, (ii) at every MCAT score level the proportion of black MCAT takers taking Step 1 is higher than for white applicants, and (iii) 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 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. Indeed we see it also in the probabilities of applying to medical school. Let us examine a few other examples to help us understand both how it occurs and what we ought to do to allow us to make sensible inferences from such results.

On September 2, 1998 the *New York Times* reported evidence of high school grade inflation. They showed that a greater proportion of high school students were getting top grades while at the same time their SAT-Math scores had declined (see Table 2). Indeed, when we look at their table, the data seem to support this claim; at every grade level SAT scores have declined by 2 to 4 points over the decade of interest. Yet the

[1] 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*."

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
Total	**8,698**	**74,213**	**82,911**	**4,891**	**43,019**	**47,910**	**2,853**	**21,636**	**24,489**	**1,966**	**19,732**	**21,698**

Table 1
(*Continued*)

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 med 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
Total	**0.56**	**0.58**	**0.58**	**0.33**	**0.29**	**0.30**	**0.23**	**0.27**	**0.26**	**0.69**	**0.91**	**0.89**

Table 2
A decade of high school grades and SAT scores: are students getting better or worse?

Grade average	Percentage of students getting grades		Average SAT Math scores		
	1988	1998	1988	1998	Change
A+	4%	7%	632	629	−3
A	11%	15%	586	582	−4
A−	13%	16%	556	554	−2
B	53%	48%	490	487	−3
C	19%	14%	431	428	−3
Overall average			**504**	**514**	**10**

From NY Times September 2, 1998.

article also reported that over the same time period (1988–1998) SAT-Math scores had in fact gone up by ten points.

How can the average SAT score increase by 10 points from 1988 to 1998, while at the same time decrease at every grade level? The key is the change in the percentages of children receiving each of the grades. Thus, although it is true that SAT-Math scores declined from 632 to 629 for A+ students, there are nearly twice as many A+ students in 1998. Thus in calculating the average score we weight the 629 by 7% in 1998 rather than by only 4%. The calculation of the average SAT score in a year partitioned by grade level requires both high school grades and SAT scores for students with those grades. As we will demonstrate shortly, we can make the anomaly disappear by holding the proportional mix fixed across the two groups.

As a third example, consider the results from the National Assessment of Educational Progress shown in Table 3. We see that 8th grade students in Nebraska scored 6 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 doesn't 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 sensi-

Table 3
NAEP 1992 8th grade Math scores

Other					
	State	White	Black	Non white	**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%	

ble 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 multiple the average score for each subgroup by their respective percentages in the nation, e.g., $(283 \times 0.69) + (242 \times 0.16) + (260 \times 0.15) = 273$. Because New Jersey's demographic mix is not very different 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 et al. (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, since 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

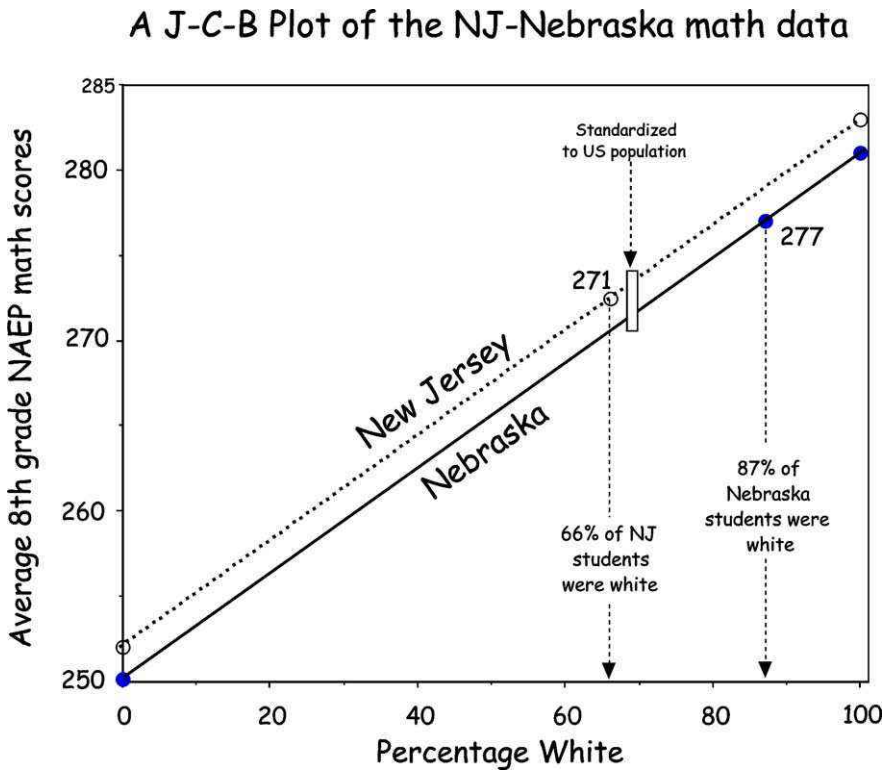

Fig. 1. A graph developed by Jeon et al. (1987) that illuminates the conditions for Simpson's Paradox as well as how standardization ameliorates it.

white students 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 4 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 to 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

Table 4
Distributions of Step 1 rates by ethnicity with standardized totals

MCAT score	Step 1 rates		Percentage of		Standardized Step 1 rates	
	Black	White	Black	White	Black	White
3 or less	1%	0%	15%	**2%**	0%	0%
4	3%	1%	14%	**3%**	0%	0%
5	10%	1%	14%	**5%**	1%	0%
6	20%	3%	15%	**7%**	1%	0%
7	31%	9%	13%	**9%**	3%	1%
8	44%	18%	14%	**16%**	7%	3%
9	48%	30%	8%	**18%**	9%	5%
10 or more	51%	43%	8%	**40%**	20%	17%
Total	23%	27%	23%		41%	27%

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.

4. Kelley's Paradox

We now turn to a second statistical paradox that appears in the same data set. The score distributions of the two ethnic groups under consideration are shown in Figure 2. The score distributions are about one standard deviation apart. This result matches rather closely what is seen in most other standardized tests (e.g., SAT, NAEP, LSAT, GRE).[2]

The distribution of MCAT scores in the medical school population is shown as Figure 3. The means of these distributions are a little closer together, but because the individuals that make up these distributions are highly selected, they have somewhat smaller standard deviations (they are leptokurtic). The difference between them, in standard deviation units, is 18% larger.

As shown by the data that make up Table 1, a black candidate with the same MCAT score as a white candidate has a greater likelihood of admission to medical school. Often much greater; at MCAT scores of 5 or 6 a black candidate's probability of admission is about twelve times that of a white candidate. There are many justifications behind a policy that boosts the chances of being selected for medical school for black applicants. One of these is the expectation that students from less privileged social backgrounds who do well must be very talented indeed and hence will do better still when given the opportunity. This point of view was more fully expressed in the August 31, 1999 issue of the *Wall Street Journal*. In it an article appeared about a research project done

[2] For further general documentation see Bowen and Bok (1998). Details on NAEP are in Johnson and Zwick (1988); on SAT see http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2003/pdf/table_3c.pdf; on GRE "*Sex, race, ethnicity and performance on the GRE General Test: 2003–2004*" (Author, 2003).

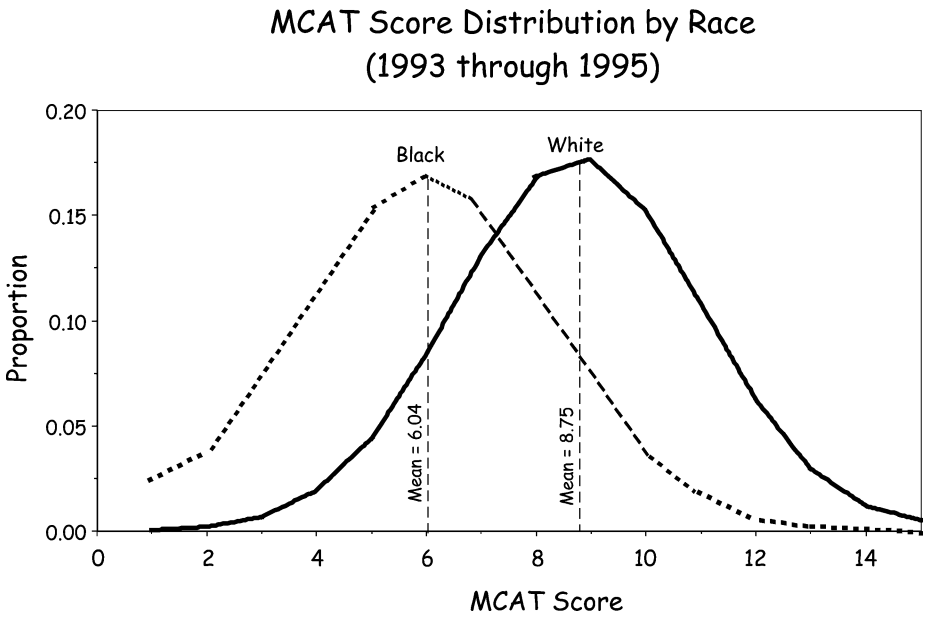

Fig. 2. MCAT score distributions by race aggregated for all examinees taking the exam for the last time between 1993 and 1995.

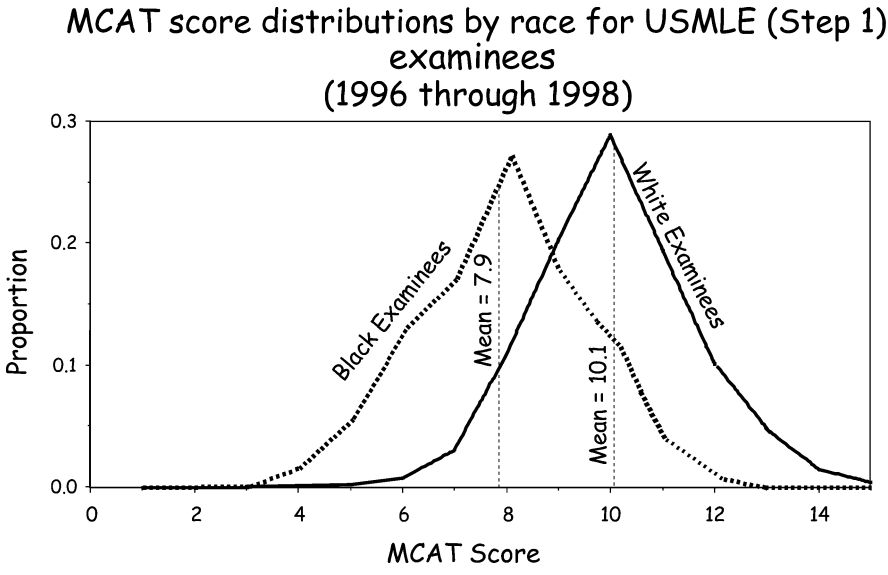

Fig. 3. MCAT score distributions by race aggregated for all medical students who took the USMLE Step 1 from 1995 through 1998.

under the auspices of the Educational Testing Service called "Strivers." The goal of "Strivers" was to aid colleges in identifying applicants (usually minority applicants) who have a better chance of succeeding in college than their test scores and high school grades might otherwise suggest. The basic idea was to predict a student's SAT score from a set of background variables (e.g., ethnicity, SES, mother's education, etc.) and characterize those students who do much better than their predicted value as "Strivers". These students might then become special targets for college admission's officers. In the newspaper interview the project's director, Anthony Carnevale said, "When you look at a Striver who gets a score of 1000, you're looking at someone who really performs at 1200". Harvard emeritus professor Nathan Glazer, in an article on Strivers in the September 27, 1999 *New Republic* indicated that he shares this point of view when he said (p. 28) "It stands to reason that a student from a materially and educationally impoverished environment who does fairly well on the SAT and better than other students who come from a similar environment is probably stronger than the unadjusted score indicates."

But before this intuitively appealing idea can be accepted it has a rather high theoretical hurdle to clear. To understand this hurdle and be able to attach its relevance to medical licensing, it is worthwhile to drop back a century and trace the origins of the fundamental issues underlying this proposal.

When we try to predict one event from another we always find that the variation in the prediction is smaller than that found in the predictor. Francis Galton (1889) pointed out that this always occurred whenever measurements were taken with imperfect precision and was what he called "regression toward the mean." This effect is seen in some historical father-child height data shown in Figure 4. Note how the father's heights vary over a 44-centimeter range (from 152 to 196 centimeters) while the prediction of their children's heights, shown by the dark dashed line, varies over only a 30-centimeter range (from 158 to 188 centimeters). What this means is that fathers that are especially tall are predicted to sire children that are tall but not as tall, and fathers who are short are predicted to have children that are short but not as short as their fathers. In this instance it is the imperfect relationship between father's and son's heights, rather than the imperfect precision of measurement, that gives rise to the regression effect.

While regression has been well understood by mathematical statisticians for more than a century, the terminology among appliers of statistical methods suggests that they either thought of it as a description of a statistical method or as only applying to biological processes. The economic statistician Frederick C. Mills (1924) wrote "the original meaning has no significance in most of its applications" (p. 394).

Stephen Stigler (1997, p. 112) pointed out that this was "a trap waiting for the unwary, who were legion." The trap has been sprung many times. One spectacular instance of a statistician getting caught was "in 1933, when a Northwestern University professor named Horace Secrist unwittingly wrote a whole book on the subject, *The Triumph of Mediocrity in Business*. In over 200 charts and tables, Secrist 'demonstrated' what he took to be an important economic phenomenon, one that likely lay at the root of the great depression: a tendency for firms to grow more mediocre over time." Secrist (1933) showed that the firms with the highest earnings a decade earlier were currently performing only a little better than average; moreover, a collection of the more poorly

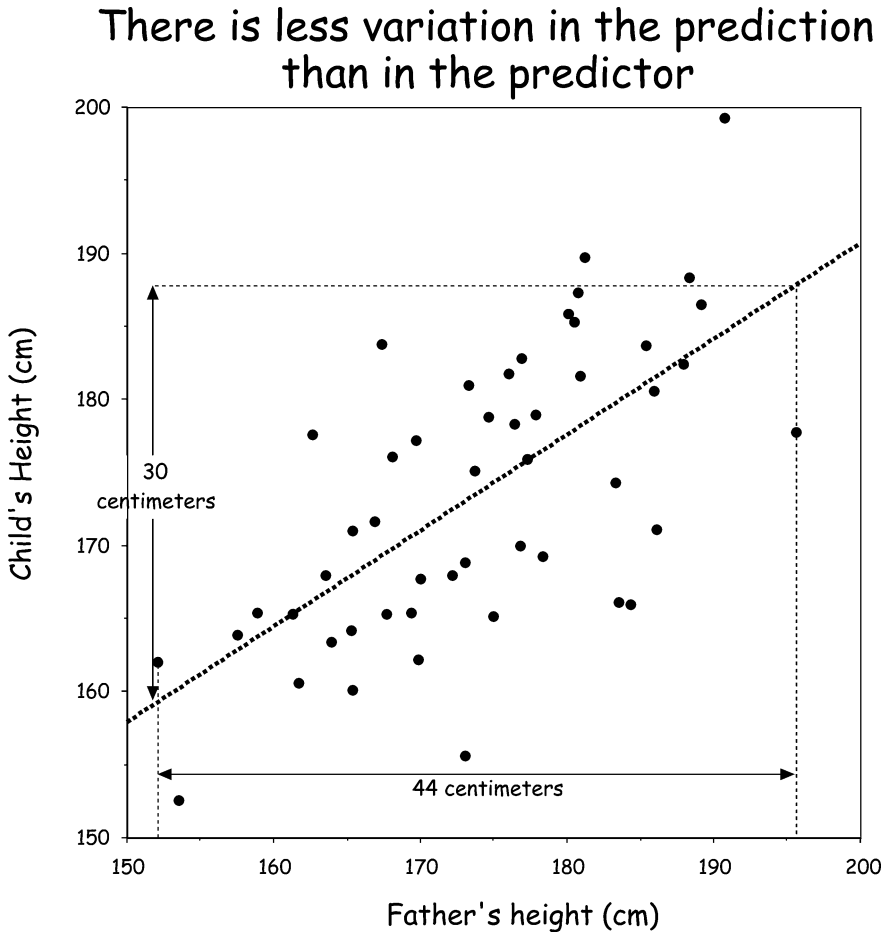

Fig. 4. The height of a child is roughly predictable from the height of its parent. Yet the distance between the shortest and tallest adults is greater than distance between the predicted statures of the most extreme children.

performing firms had improved to only slightly below average. These results formed the evidence supporting the title of the book. Harold Hotelling (1933), in a devastating review, pointed out that the seeming convergence Secrist obtained was a "statistical fallacy, resulting from the method of grouping." He concluded that Secrist's results "prove nothing more than that the ratios in question have a tendency to wander about." He then demonstrated that the firms that had the highest earnings now were, on average, only slightly above average ten years earlier. And firms with the lowest earnings now were only slightly worse, on average, than the average ten years previous. Thus showing the reverse of what Secrist claimed. Hotelling's argument was not original with him. Galton not only showed that tall fathers had sons who were, on average shorter than they were, but he also showed that very tall sons had fathers who were shorter than they were.

It is remarkable, especially considering how old and well-known regression effects are, how often these effects are mistaken for something substantive. Although Secrist himself was a professor of statistics, Willford I. King (1934) wrote a glowing review of Secrist's book, was president of the American Statistical Association! This error was repeated in a book by W.F. Sharpe (1985), a Nobel laureate in economics (p. 430) who ascribed the same regression effect Secrist described to economic forces. His explanation of the convergence, between 1966 and 1980, of the most profitable and least profitable companies was that "ultimately economic forces will force the convergence of profitability and growth rates of different firms." This continuing misunderstanding led Milton Friedman (1992), yet another Nobel laureate in economics, to try to set his colleagues straight. He discussed one of the principal theses of a book by Baumol et al. (1989), and its review by Williamson (1991); that the rates of growth of various countries tend to converge. Friedman agreed with their thesis but not their explanation, which, he pointed out, was statistical, not economic.

Truman Kelley (1927) described a specific instance of a regression formula of great importance in many fields, although it was proposed for use in educational testing. It shows how you can estimate an examinee's true score from his/her observed score on a test. "True score" is psychometric shorthand for the average of the person's observed scores if they took essentially identical tests[3] over and over again forever. Kelley's equation relates the estimated true score ($\hat{\tau}$) to the observed score (x). It tells us that the best estimate is obtained by regressing the observed score in the direction of the mean score (μ) of the group that the examinee came from. The amount of the regression is determined by the reliability (ρ) of the test. Kelley's equation is

$$\hat{\tau} = \rho(x) + (1 - \rho)\mu. \tag{1}$$

Note how Kelley's equation works. If a test is completely unreliable ($\rho = 0$), as would be the case if each examinee's score was just a random number, the observed score would not count at all and the estimated true score is merely the group mean. If the test scores were perfectly reliable ($\rho = 1$) there would be no regression effect at all and the true score would be the same as the observed score. The reliability of virtually all tests lies between these two extremes and so the estimated true score will be somewhere between the observed score and the mean.

A diagram aids intuition about how Kelley's equation works when there are multiple groups. Shown in Figure 5 are the distributions of scores for two groups of individuals, here called Group 1 (lower scoring group) and Group 2 (higher scoring group). If we observed a score x, midway between the means of the two groups the best estimate of the true score of the individual who generated that score depends on which group that person belonged to. If that person came from Group 1 we should regress the score downward; if from Group 2 we should regress it upward.

[3] "Essentially identical tests" is shorthand for what psychometricians' call "parallel forms" of the test. This means tests that are constructed of different questions but span the same areas of knowledge, are equally difficult, and are equally well put together. In fact, as one part of the formal definition is the notion that if two tests were truly parallel a potential examinee would be completely indifferent as to which form was actually presented.

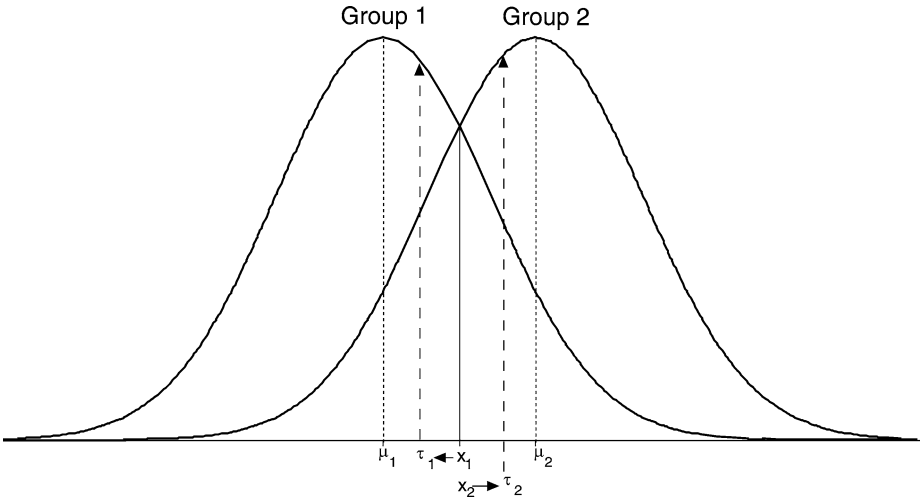

Fig. 5. A graphical depiction of Kelley's equation for two groups. The two distributions and their means are shown. Also indicated is how the true scores are regressed when two identical observed scores come from each of the two different score distributions.

The regression effect occurs because there is some error in the observed score. The average error is defined to be zero, and so some errors will be positive and some negative. Thus if someone from a low scoring group has a high score we can believe that to some extent that person is the recipient of some positive error, which is not likely to reappear upon retesting, and so we regress their score downward. Similarly, if someone from a high scoring group has an unusually low score, we regress that score upward.

So far this is merely an equation – a mathematical tautology. What is the paradox? Webster defines a paradox as a statement that is opposed to common sense and yet is true. So long as Kelley's equation deals solely with abstract groups named 1 and 2, no paradox emerges. But consider the similarity of Figure 5 with Figures 2 and 3 and consider the logic of "Strivers" which suggests explicitly that if we find someone from Group 1 with a high score, despite their coming from an environment of intellectual and material deprivation we strongly suspect their true ability ought to be considered as being somewhat higher. Similarly, someone who comes from a more privileged background, but who scores low, leads us to suspect a lack of talent and hence ought to be rated lower still. The underlying notion of "Strivers" points in the opposite direction of what would be expected through the application of Kelley's equation. This is the source of the paradox.

Harvard statistician, Alan Zaslavsky (2000) in a letter that appeared in the statistics magazine *Chance* tried to salvage the strivers idea with a more statistical argument. He did not question the validity of Kelley's equation, when it matched the situation. But he suggested that we must determine empirically what is the correct distribution toward which we regress the observed score. Does a person selected from the lower distribution remain a part of that group after being selected? Or does the very act of being selected obviate past associations? He described a metaphorical race in which

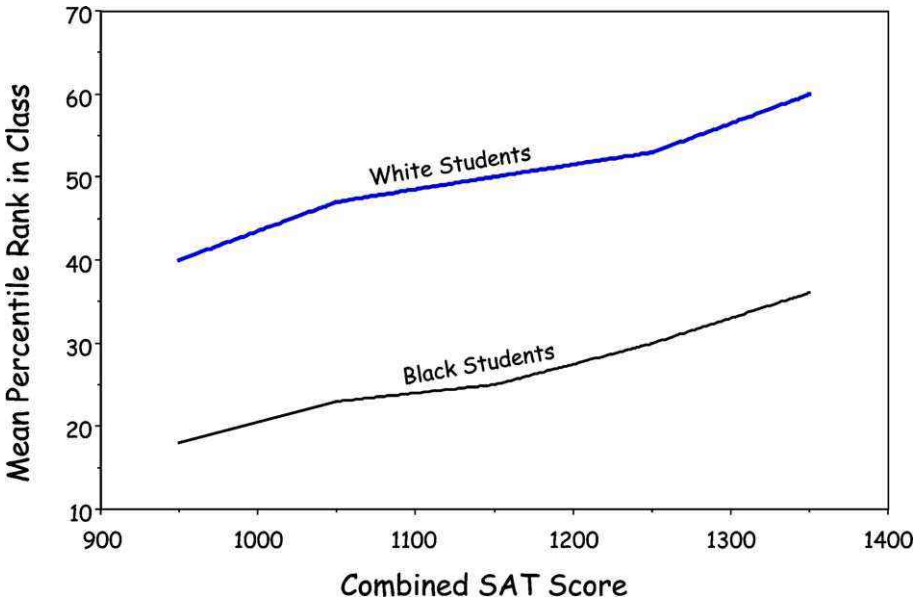

Fig. 6. An accurate revision of Figure 3.10 from Bowen and Bok (1998) showing that at all levels of SAT score Black students performance in college courses are much lower than white students with matched SAT scores. This bears out the prediction made by Kelley's equation.

one set of participants was forced to carry heavy weights. But after they were selected, they rid themselves of the weights and then ran much faster than other participants who had the same initial speed but who had not been carrying weights.

Thus the issue, as Zaslavsky views it, is an empirical one. Is social background like a weight that can be shed with proper help? Or is it more like height, a characteristic one is pretty much stuck with?

One empirical test of Kelley's equation within this context was carried out by William Bowen and Derek Bok (1998, Figure 3.10) in their exhaustive study of the value of affirmative action. They prepared a figure (analogous to Figure 6) that shows that black students' rank in college class is approximately 25 percentile points lower than white students with the same SAT scores. In this metric, the effect is essentially constant over all SAT scores, even the highest. This confirms, at least in direction, the prediction described in Figure 5; the performance in college of students with the same SAT score is different in a way predicted by their group membership.

In a more thorough analysis Ramist et al. (1994) used data from more than 46,000 students, gathered at 38 different colleges and universities to build a prediction model of college performance based on precollegiate information. One analysis (there were many others) predicted first year college grade point average from SAT score, from High School Grade Point Average (HS-GPA), and from both combined. They then recorded the extent to which each ethnic group was over or under predicted by the model. An extract of their results (from their Table 8) is shown in Table 5. The metric reported is grade points, so that a one-point difference corresponds to one grade level (a B to

Table 5
A summary of Ramist et al. (1994) results showing the size and direction of the errors in predictions for various models

	Ethnic group			
Predictor	Asian American	White	Hispanic	Black
HS-GPA	0.02	0.03	−0.24	−0.35
SAT	0.08	0.01	−0.13	−0.23
HS-GPA & SAT	0.04	0.01	−0.13	−0.16
Sample sizes	3,848	36,743	1,599	2,475

a C, for example). The entries in this table indicate the extent to which students were over-predicted by the model based on the variables indicated. Thus the entry "0.02" for Asian Americans for a prediction based on just their high school grades means that Asian Americans actually did very slightly (.02 of a grade level) better than their high school grades predicted. And the "−0.35" for Black Americans means that they did about a third of a point worse than their grades predicted. Note that while SAT scores also over-predict minority college performance, the extent of the error they make is somewhat smaller.

The results from these studies are very clear; the common regression model over-predicts non Asian minority populations. This result matches the regression phenomenon we described as Kelley's paradox. And it matches it in amount as well as direction. Moreover, these results are not the outcome of some recent social change, but have been observed for decades (Linn, 1982; Reilly, 1973).

The licensing of physicians

If Kelley's equation holds for medical students we should expect that the scores for black medical students on Step 1 should be lower than those for white students who had the same MCAT score. Figure 7 carries the principal results of our study. The error bars around each point are one standard error of the difference of the means in each direction (adjusted by the Bonferroni inequality to deal with the eight comparisons being made). Hence if the bars do not overlap the difference in the associated means is statistically significant well beyond the nominal levels.

As is evident, at each MCAT score white medical students score higher on Step 1 than matched black medical students. The shaded horizontal line stretching across the figure represents the range of passing scores used during the time period 1996 through 1998. The unusual data point for white medical students with MCAT scores of "3 or less" is not an error. It represents a single student who switched into undergraduate science courses late in her collegiate career but graduated from Wellesley with a 3.6 GPA. Her medical school gambled that despite her low MCAT score she would be a good bet. Obviously the bet paid off.

The data shown in Figure 7 match the direction predicted by Kelley's equation, but what about the amount? Figure 8 is similar to Figure 7 except that we have ungrouped

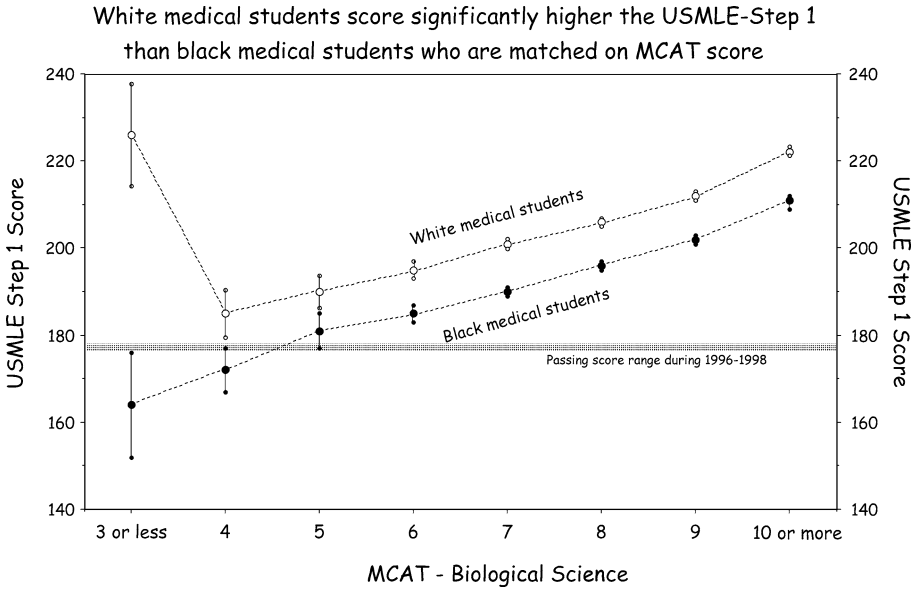

Fig. 7. Analysis of USMLE showing that at all MCAT score levels Black medical students performance on Step 1 are much lower than white medical students with matched MCAT scores. This bears out the prediction made by Kelley's equation.

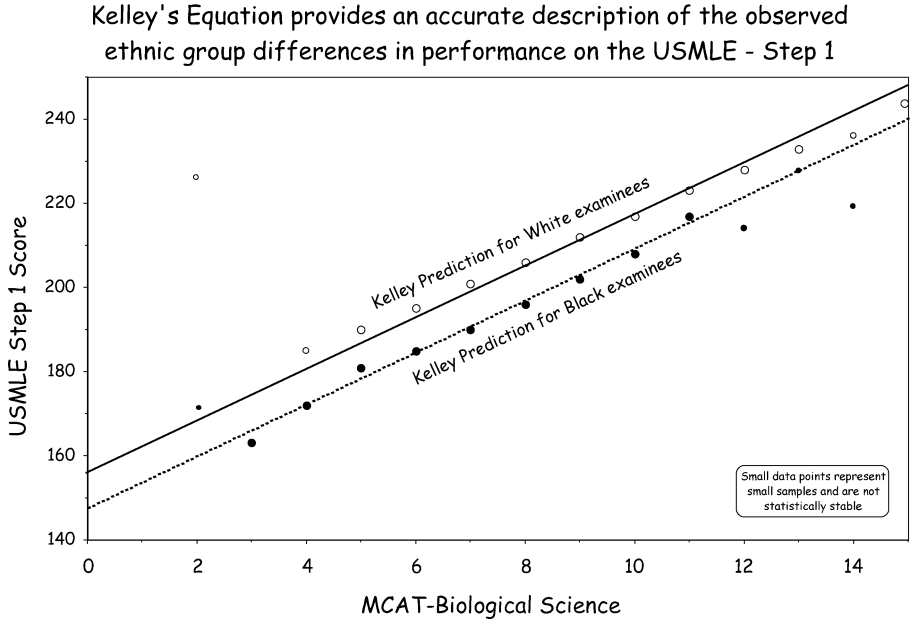

Fig. 8. A repeat of the results shown in Figure 7 with the Kelley predictions superimposed. This bears out the prediction made by Kelley's equation in both direction and size.

the data points at the ends of the distribution (some of the data points thus depicted are extremely unstable, being based on very few individuals) to present as unvarnished a view as possible. We have reduced the physical size of the data points that represent very few individuals. The lines drawn on the plot are the predictions made by substituting group means and the correlation between MCAT and USMLE into Kelley's equation. This yields two lines about ten points apart. We view this as relatively conclusive evidence that ethnicity, at least in this instance, is not like Zaslavsky's metaphorical weights, but rather it is a characteristic that is not apparently cast off when an individual enters medical school.

Unfortunately, academic ability in the instances we have examined is more like height than Zaslavsky's metaphorical weights. Neither Mr. Carnevale nor Mr. Glazer is correct. Thus when you look at a Striver who gets a score of 1000, you're probably looking at someone who really performs at 950. And, alas, a Striver is probably *weaker* than the unadjusted score indicates.

This result should be distressing for those who argue that standard admission test scores are unfair to students coming from groups whose performance on such admission tests is considerably lower than average (Freedle, 2003; Mathews, 2003) and that they under-predict the subsequent performance of such students. Exactly the opposite of that is, in fact true, for there is ample evidence that students from groups who are admitted with lower than usual credentials on average, do worse than expected.[4] The extent to which they perform more poorly is almost entirely predictable from Kelley's equation.

5. 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

[4] Due to the nested structure of the data, we also used a multi-level modeling approach to determine whether the regression effect was due to, or lessened by, differences between medical schools. We found that 10% of the variation in Step 1 scores was due to differences between medical schools and the MCAT-Biological Sciences score explained 76% of this between-school variation. Although we found significant variation in the regression effect for blacks across schools the fixed effect closely approximated the regression effect reported in this chapter. Furthermore, holding MCAT score constant, black students were predicted to have lower Step 1 scores in all 138 schools. The percentage of blacks in the sample from each school explained 18% of the variance in this effect and indicated that schools with more blacks have a smaller (albeit substantively trivial) regression effect.

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.[5]

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, and
(d) 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 paper 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 sixteen years later (Holland and Rubin, 1983) and relies heavily on Rubin's model for causal inference (Rubin, 1974).

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 9 with the bivariate dispersion within each

[5] 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.

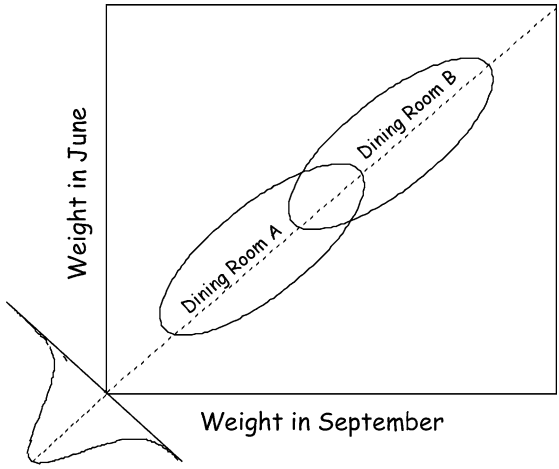

Fig. 9. A graphical depiction of Lord's Paradox showing the bivariate distribution of weights in two dining rooms at the beginning and end of each year augmented by the 45° line (the principal axis).

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 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. 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 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?" But causal questions are always comparative[6] 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 vs. what happened under the control condition.

[6] 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?"

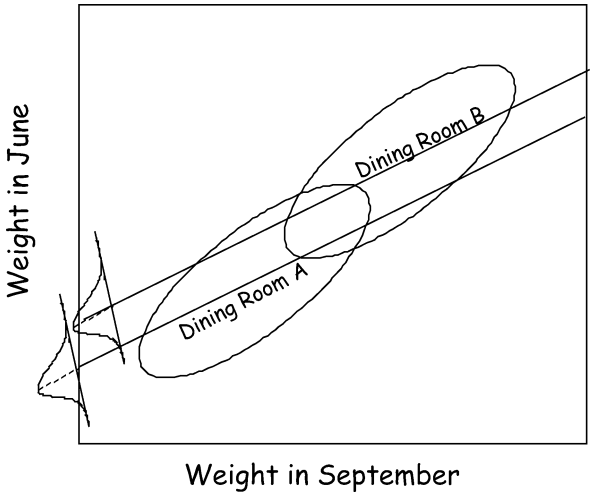

Fig. 10. A graphical depiction of Lord's Paradox showing the bivariate distribution of weights in two dining rooms at the beginning and end of each year augmented by the regression lines for each group.

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 set-up 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? To do this requires a little notation and some algebra.

The elements of the model[7] 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.
5. A concomitant variable, X.

[7] This section borrows heavily from Holland and Rubin (1983, pp. 5–8) and uses their words as well as their ideas.

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.
(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, since 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 since 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 $E(Y_C|S = C)$. Without random assignment, 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

$$E(Y_T) = E(Y_T|S = T)P(S = T) + E(Y_T|S = C)P(S = C). \tag{2}$$

There is the obvious parallel version connecting $E(Y_C|S = C)$ with $E(Y_C)$. The second term of (2) 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 Eq. (2) 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) term "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

$$D_i = E(Y_T - Y_C|G = i), \quad i = 1, 2, \tag{3}$$

and so the difference of average causal effects is

$$D = D_1 - D_2. \tag{4}$$

This can be expressed in terms of individual subpopulation averages,

$$D = \big[E(Y_T|G = 1) - E(Y_C|G = 1)\big] - \big[E(Y_T|G = 2) - E(Y_C|G = 2)\big]. \tag{5}$$

We can profitably re-arrange this to separate the observed Y_T from the unobserved Y_C

$$D = \big[E(Y_T|G = 1) - E(Y_T|G = 2)\big] - \big[E(Y_C|G = 1) - E(Y_C|G = 2)\big]. \tag{6}$$

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

$$Y_C = X \tag{7}$$

yielding,

$$E(Y_C|G = i) = E(X|G = i). \tag{8}$$

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:

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

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

$$U_i(X) = E(Y_T - X|X, G = i), \quad i = 1, 2. \tag{10}$$

Hence, the difference in these conditional ranks at X is

$$U(X) = U_1(X) - U_2(X). \tag{11}$$

The second analysis assumes that the conditional expectations in (9) are linear and parallel. Thus we can write

$$E(Y_T|X, G = i) = a_i + bX, \quad i = 1, 2. \tag{12}$$

Substituting into (10) yields

$$U_i(X) = a_i + (b - 1)X, \quad i = 1, 2. \tag{13}$$

And hence (11) simplifies to

$$U(X) = a_1 - a_2. \tag{14}$$

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 (7) that

$$Y_C = a + bX. \tag{15}$$

Where b is the common slope of the two within-groups regression lines in (12). This allows the interpretation of $U(X)$ as the difference in the causal effects D in Eq. (4).

Both of these assumptions seem to stretch the bounds of credulity, but (15) 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.

6. Conclusion

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

Will Rogers

This chapter, 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 naively, are very likely to lead us astray.

Ridding ourselves of Simpson's Paradox 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.[8]

Kelley's Paradox is not so much a summary statistic pointing in the wrong direction, as it is an indication that our intuition has been improperly trained. The fact that economists fall into this trap more often than others may reflect on their training, but why this should be the case is a phenomenon that we will not try to explain.

Lord's Paradox is the newest of this triad. 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.

Acknowledgements

We are grateful to the Association of American Medical Colleges for allowing us to use their MCAT data. More specifically we would like to 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 paper. 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 chapter was read and commented on by our colleagues Ron Nungester and Brian Clauser; we thank them for helping us excise errors and fuzzy thinking and a shortened version appeared in the *American Statistician* **58**(2), 117–123 (Wainer and Brown, 2004). We are also grateful to the sharp eyes of an anonymous referee who found several errors that had escaped our notice. 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.

[8] 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 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."

References

Author (2003). *Sex, Race, Ethnicity and Performance on the GRE General Test: 2003–2004*. Educational Testing Service, Princeton, NJ.

Baker, S.G., 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.

Baumol, W.J., Blackman, S.A.B., Wolff, E.N. (1989). *Productivity and American Leadership: The Long View*. MIT Press, Cambridge and London.

Bowen, W.G., Bok, D. (1998). *The Shape of the River*. Princeton University Press, Princeton, NJ.

Freedle, R.O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review* **73** (1), 1–43.

Friedman, M. (1992). Do old fallacies ever die? *Journal of Economic Literature* **30**, 2129–2132.

Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

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

Hotelling, H. (1933). Review of *The triumph of mediocrity in business* by Secrist H. *Journal of the American Statistical Association* **28**, 463–465.

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

Johnson, E.G., Zwick, R.J. (1988). The NAEP Technical Report, Educational Testing Service, Princeton, NJ.

Kelley, T.L. (1927). *The Interpretation of Educational Measurements*. World Book, New York.

King, W.I. (1934). Review of *The triumph of mediocrity in business* by Secrist H. *Journal of Political Economy* **42**, 398–400.

Linn, R. (1982). Ability testing: Individual differences and differential prediction. In: Wigdor, A.K., Garner, W.R. (Eds.), *Ability Testing: Uses, Consequences and Controversies, Part II*. In: *Report of the National Academy of Sciences Committee on Ability Testing*. National Academy Press, Washington, DC.

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

Mathews, J. (2003). The bias question. *The Atlantic Monthly*, 130–140.

Mills, F.C. (1924). *Statistical Methods. Applied to Economics and Business*. Henry Holt, New York.

Ramist, L., Lewis, C., McCamley-Jenkins, L. (1994). *Student Group Differences in Predicting College Grades: Sex, Language and Ethnic Group*. The College Board, New York.

Reilly, R.R. (1973). A note on minority group bias studies. *Psychological Bulletin* **80**, 130–133.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Secrist, H. (1933). *The Triumph of Mediocrity in Business*. Bureau of Business Research, Northwestern University, Evanston, IL.

Sharpe, W.F. (1985). *Investments*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

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

Stigler, S. (1997). Regression toward the mean, historically considered. *Statistical Methods in Medical Research* **6**, 103–114.

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

Wainer, H. (2000). Kelley's paradox. *Chance* **13** (1), 47–48.

Wainer, H., Brown, L. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *The American Statistician* **58**, 117–123.

Williamson, J.G. (1991). Productivity and American leadership: A review article. *Journal of Economic Literature* **29** (1), 51–68.

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

Zaslavsky, A. (2000). On Kelly's paradox. *Chance* **13** (3), 3.