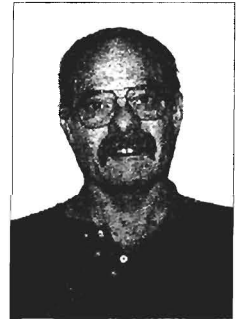


VISUAL REVELATIONS



*Howard Wainer,
Column Editor*

Simpson's Paradox

In a *Chance* article last year, Ian Westbrooke (Spring 1998, Vol. 11, No. 2, 40–42) described how the Maoris appear to be overrepresented on juries in New Zealand, 9.5% of people living within the jury districts were Maori compared with 10.1% of Maori in the pool of potential jurors. Yet, on looking more closely, Westbrooke found that “in every single local area Maori were underrepresented — often substantially.” Such anomalies are not restricted to New Zealand.

We all view the world through the filter of our own experiences. I am a statistician by training and vocation and an empiricist down to my bones. I have trouble with arguments that do not contain empirical evidence, and those that do I look at with care. The newspaper is not the same for me as it is for most other readers.

Recently I have noticed several empirical results reported in the paper, like the preceding one, that seem, at least on the surface, to be self-contradictory. When I query friends about the apparent anomalies, they snort and suggest that it must be a typo of some sort. As Westbrooke pointed out, these anomalies are examples of a common statistical artifact called Simpson's paradox (it is named after Edward Hugh Simpson, who described it in a 1951 paper, although it was recognized by the Scottish statistician George Udny Yule almost 50 years earlier).

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 1). Indeed, when we look at their table, the data seem to support this claim; at every grade level SAT scores seem to have declined by 2 to 4 points over the decade of interest. Yet in the body of the article was the information that over the time period reported (1988–1998) SAT–Math scores had in fact gone up by 10 points. How can everyone's scores go down while the average goes up?

Column Editor: Howard Wainer, Principal Research Scientist, Measurement-Statistics Data Research, Educational Testing Service (15-T), Rosedale Road, Princeton, NJ, 08541-0001, USA, E-mail hwainer@rosedale.org.

Table 1 — High School Grade Distributions and Associated SAT Math Scores During the Decade 1988–1998

Grade	% 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 *New York Times*, September 2, 1998

How does this apparently impossible result occur? The key is the change in the percentages of children receiving each of the grades. Thus although it is true that SAT–Math scores have 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 needs to use both high school grades and SAT scores for children with those grades. We can make the anomaly disappear by holding the proportional mix fixed.

This anomaly is not rare. For example, consider the results from the National Assessment of Educational Progress shown in Table 2. We see that 8th-grade students in Nebraska score 6 points higher than their counterparts in New Jersey in mathematics. Yet we also see that white students do better in New Jersey. Black students also do better in New Jersey. Indeed, all other students do better in New Jersey. How is this possible? Once again it is an example of Simpson's paradox. Because a much greater proportion of Nebraska's 8th-grade students (87%) are from the higher scoring white population than is the case in New Jersey (66%), their scores contribute more to the total.

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 the 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 based on a common demographic mixture. In this instance a 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 they had the same demographic mix as does the nation. When this is done, we find that New Jersey's score is not affected much (273 instead of 271), whereas Nebraska's scores shrinks substantially (271 instead of 277).

The third example of Simpson's Paradox was provided by Thomas D. Woolsey in his chapter in Linder and Groves (eds.) 1947 book, *Vital Statistics Rates in the United States, 1900-1940*. His data are shown in Table 3. They show that although the 1930 death rate in Maine (1,391) is higher than the death rate for the same year in South Carolina (1,289), the death rates for all age groups individually (except 5-9-year-olds) are much higher in South Carolina

Table 2 — NAEP 1992 8th-Grade Math Scores

	State	Other			Standardized
		White	Black	Nonwhite	
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%	

The answer, once again, is found by noting that there was a much higher proportion of older people in Maine, for whom the death rates are the highest. The extent of the distortion caused by the unbalanced age distributions is seen when the death rates are adjusted to reflect a common (U.S. national) age distribution: at almost any age, it's a lot safer in Maine.

Simpson's paradox can occur whenever data are aggregated. If data are collapsed across a subclassification (like grades, race, or age) the overall change observed may not represent what is going on. Standardization can help correct this, but nothing will prevent the possibility of yet another subclassification, as yet unidentified, changing things around again. But I believe that we are helped by knowing of the possibility so that we can contain the enthusiasm of our impulsive first inferences.

Simpson's paradox is well known among statisticians but is almost completely unknown among everyone else. I decided to spend this essay on something that was "old hat" in the hope that perhaps by collecting a few more examples into a single convenient place I will make it easier for teachers to spread the word.

Table 3 — Age Specific Death Rates and Populations, Maine and South Carolina, 1930

Age	Percent of population in age category		Death rate (per 100,000)		Difference
	Maine	South Carolina	Maine	South Carolina	
0-4	9.4	11.8	2,056	2,392	-336
5-9	10.0	13.9	186	185	1
10-14	9.3	12.8	140	184	-44
15-19	8.6	12.2	223	426	-203
20-24	7.6	9.6	370	645	-275
25-34	13.3	12.6	391	871	-480
35-44	12.7	11.0	545	1,242	-697
45-54	11.3	8.3	1,085	1,994	-909
55-64	9.1	4.6	2,036	3,313	-1,277
65-74	5.8	2.3	5,219	6,147	-928
75+	2.8	1.0	13,645	14,136	-491
Total death rate (per 100,000 population)			1,391	1,289	102
Adjusted death rate (per 100,000 population)			1,203	1,716	-513

References and Further Reading

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

Westbrooke, F. (1998), "Simpson's Paradox: An Example in a New Zealand Survey of Jury Composition," *Chance*, 11(2), 40-42

Woolsey, T. D. (1947), "Adjusted Death Rates and Other Indices of Mortality," Chapter 4 in *Vital Statistics Rates in the United States, 1900-1940*, eds. F.E. Linder and R.D. Grove, Washington, DC: National Office of Vital Statistics, U.S. Government Printing Office, pp. 60-91.

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