Instances of Simpson's Paradox

Thomas R. Knapp, University of Rochester, Rochester, NY

In its simplest form Simpson's Paradox goes like this:

> Consider two populations* for which the overall rate r of occurrence of some phenomenon in Population A is greater than the corresponding rate R in Population B. Suppose that each of the two populations is composed of the same two categories C_1 and C_2 and the rates of occurrence of the phenomenon for the two categories in Population A are r_1 and r_2, and the rates of occurrence in Population B are R_1 and R_2. If $r_1 < R_1$ and $r_2 < R_2$, despite the fact that $r > R$, then Simpson's Paradox is said to have occurred.

The following fictitious example of the batting averages of two baseball players illustrates the paradox rather dramatically:

	Player A			Player B		
	Times at bat	Hits	Average	Times at bat	Hits	Average
Against right-handed pitchers (C_1)	202	45	.223($= r_1$)	250	58	.232($= R_1$)
Against left-handed pitchers (C_2)	250	71	.284($= r_2$)	108	32	.296($= R_2$)
Overall	452	116	.257($= r$)	358	90	.251($= R$)

How can this be? How can Player A have a better overall batting average than Player B yet be worse against both right-handed pitchers and left-handed pitchers? Answer: He can, if the two players don't bat against right-handed pitchers the same percentage of the time.

The key to the paradox is the differential weighting of the rates for C_1 and C_2 when the overall rate for each population is computed. Since $r = w_1r_1 + w_2r_2$ (where $w_1 + w_2 = 1$) and $R = W_1R_1 + W_2R_2$ (where $W_1 + W_2 = 1$ but where W_1 and W_2 may be very different from w_1 and w_2), there is no necessary ordering of r and R no matter what r_1, r_2, R_1, and R_2 are. In the artificial example just cited, Player A batted against right-handed pitchers 202 out of 452 times at bat, or 44.7% of the time ($w_1 = .447$ and $w_2 = .553$), whereas Player B batted against right-handed pitchers 250 out of 358 times at bat, or 69.8% of the time ($W_1 = .698$ and $W_2 = .302$). This discrepancy in relative weights was large enough to switch the rank-order of the batting averages of the two players.

So Simpson's Paradox *can* occur. The more interesting matter is whether or not it *does* occur. Yes it does, although apparently not very often. C.H. Wagner cites three

*The "populations" can be people, time periods, etc. as well as cities, states, and the like, which are populations in the demographic sense. There is also nothing special about *two* populations or *two* categories, but this simplest case is the one which E. H. Simpson discussed (The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society*, Series B, 13 [1951] 238–241). He also dealt primarily with probabilities rather than rates, but the paradox can be couched in terms of rates, probabilities, percentages, proportions, ratios, or averages.

instances ("Simpson's Paradox in Real Life," *The American Statistician*, 36 [1982] 46–48):

1. The overall subscription renewal rate for *American History Illustrated* magazine increased from January, 1979 to February, 1979 but the rate decreased for each category of subscriber.
2. The overall federal income tax rate increased from 1974 to 1978 but decreased for each income bracket.
3. The overall death rate from tuberculosis in 1910 was greater in Richmond, Virginia than in New York City but was less for whites and less for non-whites.

In my studies of the seasonality of births, I have found several other instances. In the United States, births are relatively heavy in the late summer and early fall, and relatively light in the spring. This phenomenon is especially pronounced in the southern states for non-whites. A simple measure of seasonality is the ratio S/A of September to April births (both months have 30 days). That ratio has consistently been greater than 1 in almost every part of the country for the 50 or so years that good birth records have been kept.

I have been able to obtain from the National Center for Health Statistics 15 computer tapes containing over 2.5 million birth records for 1976. In calculating the overall S/A ratios in 1976 for the various states, and for whites (W) and non-whites (NW) separately, I uncovered the following data:

	Alabama				Texas				Georgia			
	A	S	$A+S$	S/A	A	S	$A+S$	S/A	A	S	$A+S$	S/A
W	2894	3240	6314	1.182	14,052	16,926	30,978	1.205	3878	4548	8426	1.173
NW	1426	1954	3380	1.370	2,152	3,010	5,162	1.399	1896	2622	4518	1.383
Overall	4320	5374	9694	1.244	16,204	19,936	36,140	1.230	5774	7170	12,944	1.242

The paradox occurs in both the Alabama vs. Texas and the Texas vs. Georgia comparisons, and is attributable to the fact that white births constituted a smaller percentage of all births in both Alabama (6314 out of 9694, or 65.1%) and Georgia (8426 out of 12,944, which is also 65.1%) than in Texas (30,978 out of 36,140, or 85.7%). The troublesome thing about these data is that there is no easy answer to the question, "Which of the two states (Alabama vs. Texas or Texas vs. Georgia) had the greater seasonality?" Looking at the data one way (separately by race), Alabama and Georgia were both less seasonal than Texas, but looking at the data another way (by overall totals), Texas is the less seasonal. (Note that the Alabama vs. Georgia comparison is a bit complicated, but the interpretation is unambiguous: white births were slightly more seasonal in Alabama, and non-white births were slightly more seasonal in Georgia.)

Birth data for other years yielded a few more instances of Simpson's Paradox. In 1968 and in 1972, the District of Columbia had a greater overall S/A ratio than Virginia, but smaller S/A ratios for both whites and non-whites. White births constituted about 15% of all births in the District of Columbia and about 75% of all births in Virginia, for those years.

Simpson's Paradox is a non-intuitive phenomenon which is relatively unknown and seldom taught. For some additional information about this paradox and related matters, the interested reader is referred to the following sources:

REFERENCES

1. M. R. Cohen and E. Nagel, An Introduction to Logic and Scientific Method, Harcourt, Brace and Company, New York, 1934, p. 449.
2. R. Falk and M. Bar-Hillel, Magic possibilities of the weighted average, Mathematics Magazine 53 (1980) 106–107.
3. D. Freedman, R. Pisani, and R. Purves, Statistics, Norton, New York, 1978, pp. 12–15.
4. H. E. Reinhardt, Some statistical paradoxes, in Shulte, A.P. and Smart, J.R. (eds), Teaching Statistics and Probability, Yearbook of the National Council of Teachers of Mathematics, Reston, Va., 1981.
5. C. H. Wagner, Simpson's paradox, Undergraduate Mathematics and its Applications (UMAP) Journal 3 (1982) 185–198.

Approximating Solutions for Exponential Equations

Norman Schaumberger, Bronx Community College, Bronx, NY

To find x in $2^x = 5$, everyone will suggest, "use logarithms." But we are after something different. We would like to improve students' ability to manipulate exponents and increase their skills in estimation by obtaining x without the formal use of logarithms.

We begin by observing that $2^7 = 128$ is approximately equal to $5^3 = 125$. Thus, we write

$$5^3 \approx 2^7 \quad \text{or} \quad 5 \approx 2^{7/3} = 2^{2.333\ldots}$$

Since $2^{2.3219281\ldots} = 5$, our estimate is quite good (missing by only one-half of a percent.)

Next we seek relations that can be used to estimate 3, 7, and 11 each as powers of 2. We need only be concerned about prime integers since composites are products of primes. Thus, we want to approximate primes p as $p = 2^x$ for rational x.

For $p = 2$, clearly $x = 1$.

If $p = 3$, then $3^2 \approx 2^3$ yields

$$3 \approx 2^{3/2} \text{ (via logs, } 3 = 2^{1.5849625\ldots}\text{)}.$$

If $p = 7$, then $7^2 \approx 48 = 2^4 \cdot 3 \approx 2^4 \cdot 2^{3/2} = 2^{11/2}$ yields

$$7 \approx 2^{11/4} \text{ (via logs, } 7 = 2^{2.8073549}\text{)}.$$

If $p = 11$, then $11^2 \approx 120 = 2^3 \cdot 3 \cdot 5 \approx 2^3 \cdot 2^{3/2} \cdot 2^{7/3} = 2^{41/6}$ yields

$$11 \approx 2^{41/12} \text{ (via logs, } 11 = 2^{3.4594316\ldots}\text{)}.$$

In general, $p^2 \approx p^2 - 1 = (p - 1)(p + 1)$. Since $p - 1$ and $p + 1$ are composite, both factor into a product of primes all of which are less than p. Therefore, using the preceding approximations for each prime less than p, we obtain an approximation for p itself.

This algorithm may or may not produce better approximations for p than other generating relations. For example: