Simpson's Paradox and Major League Baseball's Hall of Fame

by

Steven M. Day
Ohio State University - Agricultural Technical Institute
Wooster OH 44691

Steven M. Day earned a Master of Arts in Mathematics and a Master of Arts in Teaching Mathematics at the University of California at Davis in 1990. He has taught technical mathematics, statistics and computer applications at the Ohio State University Agricultural Technical Institute in Wooster, Ohio since 1990.

Jim and Ron are two former Major League Baseball players who played for the same team during two consecutive World Series. For the first Series, Jim's batting average was .400 while Ron's was .379. For the following Series, Jim's batting average was .167 while Ron's was .143. Thus Jim apparently outperformed his teammate, Ron, at batting for each of the two World Series. Yet, when hits and at bats for the two Series were tallied, Ron had the better overall batting average: .333 to .273 (table 1). How can this be?

Table 1. Simpson's Paradox: The Reversal Occurs when Real Data from Two Consecutive World Series Are Combined.

	Series 1			Series 2			Combined		
	AB	H	Avg.	AB	H	Avg.	AB	H	Avg.
Ron	29	11	.379	7	1	.143	36	12	.333
Jim	10	4	.400	12	2	.167	22	6	.273

AB = At Bats, H = Hits.

This is an example of what Blyth (1972) called Simpson's Paradox. Simpson (1951) described, "the dangers of amalgamating 2x2 tables," and pointed out that statisticians had long been aware of these dangers. In its statistical form, the "paradox" can occur in virtually any two by two stratification of data. Simpson (1951) gave a wonderful, hypothetical example involving a baby who separated an ordinary deck of playing cards into two groups; in each group, the ratio of red face cards to face cards was lower than the ratio of red plain cards to plain cards, while in the deck as a whole those ratios were, of course, equal (table 2). Blyth (1972) gave a hypothetical example involving rates of recovery of two groups of patients given two different medications; medication A seemed in each group to give the better recovery rate, but when the data from the two groups were combined, medication B was better (table 3). Cohen (1986) gave hypothetical and real life

Table 2. Simpson's Playing Card Example.

	Group 1		Group 2		Combined	
	Red	Black	Red	Black	Red	Black
Face	4	3	2	3	6	6
Plain	8	5	12	15	20	20

Note that 4:7 < 8:13 and 2:5 < 12:27 but 6:12 = 20:40.

examples involving age specific rates of mortality in two different countries; for each age group, country A had the higher mortality rate, but when all ages were combined, country B's mortality rate came out higher (table 4). Mitchem (1989) and Beckenbach (1979) gave hypothetical examples involving batting averages of baseball players for two seasons. Other examples, real and hypothetical, may be found in Shapiro (1982), Paik (1985), and Wagner (1982).

Table 3. Blythe's Medication Example.

	Group 1		Group 2		Combined	
	R	D	R	D	R	D
Treatment A	1000	10000	100	9	1100	10009
Treatment B	9	100	10000	1000	10009	1100

R = Recovered, D = Died. For each group, treatment A has a better recovery rate, yet overall, treatment B is far superior.

In this paper, we review the mathematics of Simpson's Paradox and use a graphical method for constructing examples of it which was given by Goddard (1991) to examine the frequency with which the paradox can occur in a special, symmetrical case. Next we uncover some real life examples of the paradox from the annals of Major League Baseball. Finally we pose some problems which, upon reflection, will be seen to involve this paradox directly.

Table 4. Cohen's Hypothetical Death Rate Example

	Young		Old		Combined	
	At Risk	Deaths	At Risk	Deaths	At Risk	Deaths
Country A	90	25	10	4	100	29
Country B	40	10	60	20	100	30

Unravelling the Paradox

Mathematically, Simpson's Paradox consists of real numbers, a_i, m_i, b_i, n_i, i = 1 to k, with the following properties:

$$0 \le a_i \le m_i, 0 \le b_i \le n_i$$

$$\frac{a_i}{m_i} > \frac{b_i}{n_i} \text{ for } i = 1 \text{ to } k \qquad (1)$$

but,

$$\frac{a}{m} \le \frac{b}{n}$$

where letters without subscripts represent sums over i of corresponding letters with subscripts.

Restricting attention to the case $k = 2$, we may get an idea of what is happening by considering a somewhat extreme example.

Table 5. A Somewhat Extreme Example of the Symmetric Case.

	Season 1			Season 2			Combined		
	AB	H	Avg.	AB	H	Avg.	AB	H	Avg.
Player A	100	60	.600	900	100	.111	1000	160	.160
Player B	900	450	.500	100	10	.100	1000	460	.460

In table 5, player B's average of .500 for the first season is not as good as player A's .600, but it is weighted much more heavily. A necessary condition for player B to come out on top in the aggregate average is for one of B's season averages to be better than player A's aggregate average; then, if that season is weighted heavily enough, B's aggregate average can be higher than A's.

The reversal of inequalities in the paradox can be extreme. As Blyth (1972) pointed out, it is possible to have

$$\frac{a_1}{m_1} > \frac{b_1}{n_1} = 0$$

and

$$1 = \frac{a_2}{m_2} > \frac{b_2}{n_2}$$

but

$$0 = \frac{a}{m} < \frac{b}{n} = 1.$$

To see how these extremes might be approached, consider that for any positive integer N we have

$$\frac{N}{N^2} > \frac{1}{N+1} \sim 0$$

and

$$1 \sim \frac{N}{N+1} > \frac{N^2-N}{N^2}$$

but

$$0 \sim \frac{2N}{N^2+N+1} < \frac{N^2 \cdot N+1}{N^2+N+1} \sim 1.$$

With sufficiently large N, we can make the approximations above as precise as we wish. In fact, if we allow $b_1 = 0$ and $a_2 = m_2$, the first two sets of approximations can be replaced with equalities.

$$\frac{1}{N} > \frac{0}{1} = 0$$

and

$$1 = \frac{1}{1} > \frac{N-1}{N}$$

but

$$0 \sim \frac{2}{N+1} < \frac{N-1}{N+1} \sim 1.$$

Note that in the foregoing analysis, $m_1 = n_2$ and $m_2 = n_1$. This is also the case in table 5 where $m_1 = n_2 = 100$, $m_2 = n_1 = 900$ and the total number of at bats is the same for each player. This symmetrical case is especially intriguing, and we shall consider it in some detail shortly.

The Poverty of Examples

Although the literature is full of hypothetical and real life examples of Simpson's Paradox, it is certainly true that one's intuition is correct far more often than not; in general, Simpson's Paradox is quite rare. Just how rare it is may be explored by way of a picture given in Goddard (1991). We develop that picture, then consider the frequency of examples of the paradox, and consider, for a special case, extreme values of that frequency.

Fixing $k = 2$ in (1) and rearranging some of the inequalities there we obtain

$$a_1 > \frac{b_1 m_1}{n_1}, \qquad a_2 > \frac{b_2 m_2}{n_2}$$

and (2)

$$a = a_1 + a_2 \leq \frac{bm}{n}.$$

Fixing b_1, b_2, n_1, n_2, m_1 *and* m_2, we allow a_1 and a_2 to vary and graph the three lines

$$a_1 = \frac{m_1 b_1}{n_1}\ ,\ a_2 = \frac{m_2 b_2}{n_2}$$
$$a_1 + a_2 = \frac{mb}{n} \tag{3}$$

obtaining a triangle of pairs (a_1,a_2) defined by (2) which yield examples of the paradox (figure 1).

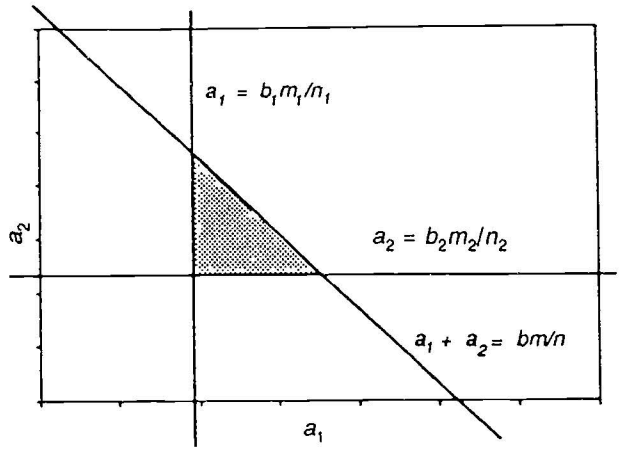

Figure 1: Goddard's Triangle. Here, all parameters are fixed except a_1 and a_2.

Under the assumptions (1) we have a rectangle from which a_1 and a_2 must come (A player, for example, cannot have more hits than "at bats")

$$0 \le a_1 \le m_1,\ 0 \le a_2 \le m_2. \tag{4}$$

Two interesting questions present themselves. First, how large can the triangle of figure 1 be? In general, there is no limit to how large it can be, but if we fix some of the parameters and allow others to vary, then the size is limited. The second, more interesting, question concerns the ratio of the area of the triangle ((2), figure 1) to the area of the rectangle (4). How great can this ratio be? There is a complication to consider and many parameters. We will consider these two questions in detail for a special, symmetrical case wherein $m_1 = n_2$ and $m_2 = n_1$.

In the case $m_1 = n_2$, $m_2 = n_1$, our equations (3) become

$$a_1 = \frac{b_1 n_2}{n_1} \qquad a_2 = \frac{b_2 n_1}{n_2} \tag{5}$$
$$a_1 + a_2 = b_1 + b_2.$$

To have any triangle (2), the diagonal line in figure 1 must lie above the intersection of the other two lines. This means we must have

$$\frac{b_1 n_2}{n_1} + \frac{b_2 n_1}{n_2} < b_1 + b_2. \tag{6}$$

To produce an example of figure 1, we fix all parameters except a_1 and a_2. If we now allow $m_1 = n_2$ to vary, the area of the triangle of figure 1 will vary also, and subject to (6) we can obtain bounds on n_2 between which the triangle will have some area. If we multiply the inequality (6) by n_2 and rearrange the terms, we obtain

$$b_1 n_2^2 - (b_1 + b_2) n_1 n_2 + b_2 n_1^2 < 0$$

which is quadratic in n_2. The solutions to the corresponding quadratic equation are

$$n_2 = n_1 \text{ and } n_2 = \frac{b_2 n_1}{b_1}.$$

Hence if we assume that $b_2 > b_1$, inequality (6) implies

$$n_1 \le n_2 \le \frac{b_2 n_1}{b_1}. \tag{7}$$

How big can the triangle be? Its base is equal to its height and equal to

$$b_1 + b_2 - \frac{b_1 n_2}{n_1} - \frac{b_2 n_1}{n_2}.$$

The area is then

$$A = \frac{1}{2}\left[b_1 + b_2 - \frac{b_1 n_2}{n_1} - \frac{b_2 n_1}{n_2}\right]^2.$$

Within the limits on n_2 given by (7), the partial derivative of this area function, A, with respect to n_2 is zero and the area of the triangle is maximized when

$$n_2 = n_1 \sqrt{\frac{b_2}{b_1}}.$$

Note that this is the geometric mean of the limits on n_2 in (7).

The question of the ratio of the area of the triangle in figure 1 to that of the rectangle (4) is complicated by the fact that part of the triangle may lie outside the rectangle. To avoid this complication and ensure that the triangle will lie entirely within the rectangle throughout the range (7) of values of n_2 it is sufficient to require that

$$b_2 \le n_1. \tag{8}$$

We ask, then, what is the maximum value of

$$\frac{\frac{1}{2}\left[b_1 + b_2 - \frac{b_1 n_2}{n_1} - \frac{b_2 n_1}{n_2}\right]^2}{n_1 n_2} \tag{9}$$

as n_2 ranges between the limits (7)? (This ratio is not correct if part of the triangle lies outside the rectangle, of course; a suitable portion of the triangle's area would have to be subtracted in that case. We require (8), and thus avoid this issue.)

Taking the partial derivative of (9) with respect to n_2 and setting it equal to zero leads one to seek solutions to a fourth degree polynomial in n_2. Fortunately, two of the solutions are the limits on n_2 imposed by (7) and we can divide the fourth degree equation by

$$(n_2 - n_1)\left(n_2 - \frac{b_2 n_1}{b_1}\right)$$

to obtain a quadratic equation in n_2. The positive solution to the resulting equation is

$$n_2 = \frac{-n_1(b_1 + b_2) + n_1\sqrt{(b_1 + b_2)^2 + 12b_1 b_2}}{2b_1}. \tag{10}$$

As an example, consider $n_1 = 200$ and $b_2 = 180$ (figure 2). The values given by (10) are then $n_{2\max} = 225.88$, 233.79 and 242.44 for $b_1 = 140$, 130 and 120 respectively, with corresponding ratios .00006958, .0001756, and .0003769. The maximum value of the ratio increases as b_1 approaches zero and/or as b_2 approaches n_1. For the special, symmetrical case under consideration, wherein we assume (8) so that the triangle lies entirely within the rectangle (4) throughout the limits (7), the limit of the ratio (9) is 2/27. For this symmetric case, if we use the ratio of the area of the triangle in figure 1 to that of the area of the rectangle of all allowable pairs (a_1,a_2) as a measure of the likelihood of the occurrence of Simpson's Paradox, we see that that likelihood is quite small.

Major League Baseball's Hall of Fame

The paradox is rare. Major League Baseball has a rich and varied history, though, and it is a history brimming with statistics. With diligent searching, examples can be found. Former major leaguers Jim and Ron of the introduction are Jim Lefebvre and Ron Fairly, who played for the Los Angeles Dodgers in the 1965 and 1966 World Series. I began my search for Simpson's Paradox with World Series because several players' statistics are conveniently displayed together in that section of Reichler (1988). Soon, however, I turned my attention to full seasons of some of the greatest baseball players of all time. Here are three notable discoveries which the reader may easily verify by consulting Reichler (1988 or later edition).

In each of his first three years in major league baseball, upstart Lou Gehrig,

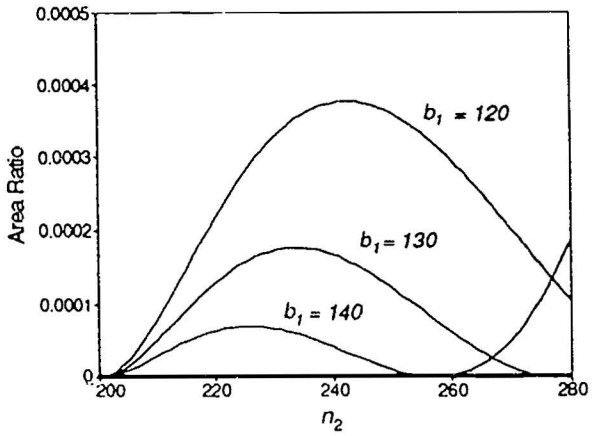

Figure 2: The ratio of Goddard's triangle to the rectangle of possible pairs (a_1, a_2) as a function of n_2. Here, $n_1 = m_2 = 200$ and $b_2 = 180$.

playing for the New York Yankees, had a better batting average than his teammate, the veteran Babe Ruth. When the data for the three years were combined, however, Ruth's average was superior, perhaps pointing out to the young Gehrig that to beat the Sultan of Swat, it wouldn't be enough just to beat him one year at a time (table 6).

Table 6. Yankee Teammates Ruth and Gehrig, 1923–1925.

	1923		1924		1925		Combined	
	AB	H	AB	H	AB	H	AB	H
Ruth	522	205	529	200	359	104	1410	509
Gehrig	26	11	12	6	437	129	475	146

Table 7. Hall of Famers Babe Ruth and Rogers Hornsby.

	1934		1935		Combined	
	AB	H	AB	H	AB	H
Ruth	365	105	72	13	437	118
Hornsby	23	7	24	5	47	12

Ruth was at it again in 1934 and 1935. Fellow Hall of Famer Rogers Hornsby beat him in each of those years, but again Ruth was better when the years were combined (table 7).

In 1941 and 1942, Stan Musial did battle with Joe DiMaggio. Stan "The Man" came out on top each year, but "Joltin'" Joe's average was better for the two years combined (table 8).

Table 8. Joe DiMaggio and Stan Musial.

	1941		1942		Combined	
	AB	H	AB	H	AB	H
Musial	47	20	467	147	514	167
DiMaggio	541	193	610	186	1151	379

The Paradox in Other Forms

We conclude with some problems. The first is simply the paradox as defined in (1) with $k = 50$. The others illustrate how the paradox may arise in fields other than statistics.

1. Two basketball players, Grace and Asina, play the same position on the same team. They each play every game, though not necessarily the same number of minutes. They play 50 games and at the end of the 50th game each has taken the same total number of shots over the course of the 50 games. Grace has a better shooting average than Asina during every one of the games, but Asina has the better average over the entire 50 games combined. How can this happen?

2. Let $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{b}_1$, and $\boldsymbol{b}_2$ be vectors in standard position each with tip in quadrant I. The angles of these vectors are α_1, α_2, β_1 and β_2 respectively. If $\alpha_1 > \beta_1$ and $\alpha_2 > \beta_2$, then how does the angle of $\boldsymbol{a}_1 + \boldsymbol{a}_2$ compare with that of $\boldsymbol{b}_1 + \boldsymbol{b}_2$?

3. Let z_1, z_2, w_1 and w_2 be numbers in the complex plane with real and imaginary parts all positive. If $\arg(z_1) > \arg(w_1)$ and $\arg(z_2) > \arg(w_2)$, then how does $\arg(z_1 + z_2)$ compare with $\arg(w_1 + w_2)$?

4. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two matrices with entries from $\mathbb{R}^+$ such that $\det\boldsymbol{A} > 0$ and $\det\boldsymbol{B} > 0$, what can be said about $\det(\boldsymbol{A} + \boldsymbol{B})$?

5. Consider the following "ordering" on $\mathbb{R}^+ \times \mathbb{R}^+$:

$$Def: \quad (a,b) <^* (c,d) \quad if \quad \frac{b}{a} < \frac{d}{c}.$$

What properties does this ordering have? Assuming addition is defined in the usual way for vectors, how does this ordering relate to the addition?

Summary

It seems reasonable to surmise that if a hypothesis is supported by two independent trials, then it will be supported when the data from those trials are combined. We have seen how Simpson's Paradox confounds this idea. It does so rarely, but in any forum, from playing cards to Major League Baseball's Hall of Fame.

References

Beckenbach, E. F. (1979). Baseball statistics. *The Mathematics Teacher,* 72, 351-352.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association,* 67, 364-366.

Cohen, J. E. (1986). An uncertainty principle in demography and the unisex issue. *The American Statistician,* 40, 32-39.

Gardner, M. (1976). On the fabric of inductive logic and some probability paradoxes. *Scientific American,* 234, 119-124.

Goddard, M. J. (1991). Constructing some categorical anomalies. *The American Statistician,* 45, 129-134.

Mitchem, J. (1989). Paradoxes in averages. *The Mathematics Teacher,* 82, 250-253.

Paik, M. (1985). A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *The American Statistician,* 39, 53-54.

Reichler, J. L. (Ed.). (1988). *The Baseball Encyclopedia* (7th ed.). New York: MacMillan.

Shapiro, S. H. (1982). Collapsing contingency tables - A geometric approach. *The American Statistician,* 36, 43-46.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society,* Ser. B, 13, 238-241.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician,* 36, 46-47.

Lucky Larry # 11

While solving a radical equation, Lucky Larry used the student method of "squaring parts." For him, of course, it worked.

$$\sqrt{x+3} + \sqrt{2x+7} = 1$$

$$\left(\sqrt{x+3}\right)^2 + \left(\sqrt{2x+7}\right)^2 = 1^2$$

$$x + 3 + 2x + 7 = 1$$

$$3x = -9$$

$x = -3$, which checks!

Submitted by Harriett Beggs
SUNY College of Technology
Canton NY 13617